

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Hypothesis Testing

(Multiple Testing)

Hypothesis Testing: (Multiple testing)

- Issues with multiple testing:

- Suppose we run m hypothesis tests on the data with null hypotheses $H_{01}, \dots, H_{0m}$
- Naïve method**: For every H_{0i} , it is rejected if the p-value is below the significance level α .

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- The **Family Wise Error Rate (FWER)** is defined as:

$$FWER = \Pr[V \geq 1]$$

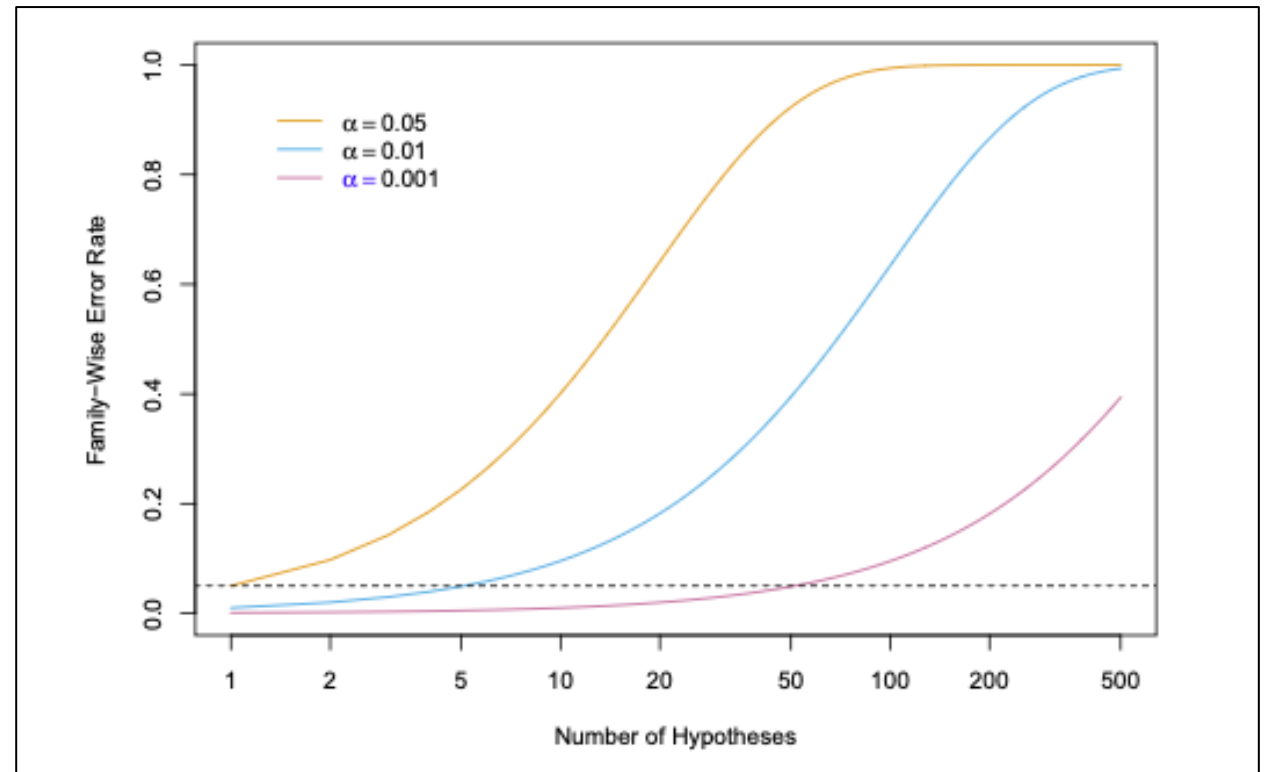
Hypothesis Testing: (Multiple testing)

- Issues with multiple testing:

- Suppose we run m hypothesis tests on the data with null hypotheses $H_{01}, \dots, H_{0m}$
- Naïve method**: For every H_{0i} , it is rejected if the p-value is below the significance level α .

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- The **Family Wise Error Rate (FWER)** is defined as:
$$FWER(\alpha) = \Pr[V \geq 1] = 1 - \Pr[V = 0] = 1 - (1 - \alpha)^m \text{ (independence and all null)}$$



Hypothesis Testing: (Multiple testing)

- Issues with multiple testing:
 - Suppose we run m hypothesis tests on the data with null hypotheses $H_{01}, \dots, H_{0m}$
 - **Naïve method**: For every H_{0i} , it is rejected if the p-value is below the significance level α .
- **Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$
- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?

Hypothesis Testing: (Multiple testing)

- Issues with multiple testing:
 - Suppose we run m hypothesis tests on the data with null hypotheses $H_{01}, \dots, H_{0m}$
 - **Naïve method**: For every H_{0i} , it is rejected if the p-value is below the significance level α .
- **Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$
- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?
$$\begin{aligned} FWER(\alpha) &= \Pr[\text{at least one null hypothesis is falsely rejected}] \\ &\leq \sum_j \Pr[H_{0j} \text{ is falsely rejected}] \quad (\text{from union bound}) \\ &= m \cdot \alpha \end{aligned}$$
- Since we want $m \cdot \alpha \leq \alpha'$, we can use $\alpha = \alpha' / m$. This is known as the **Bonferroni method**.

Hypothesis Testing: (Multiple testing)

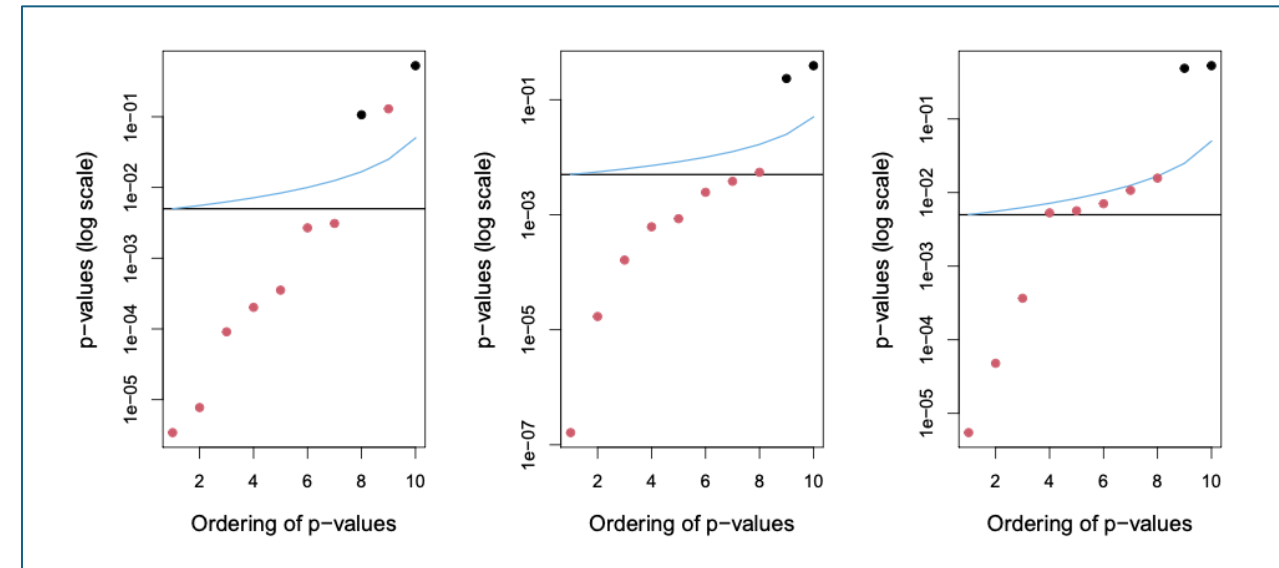
- Issues with multiple testing:
 - Suppose we run m hypothesis tests on the data with null hypotheses $H_{01}, \dots, H_{0m}$
 - **Naïve method**: For every H_{0i} , it is rejected if the p-value is below the significance level α .
- **Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$
- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?
$$\begin{aligned} FWER(\alpha) &= \Pr[\text{at least one null hypothesis is falsely rejected}] \\ &\leq \sum_j \Pr[H_{0j} \text{ is falsely rejected}] \quad (\text{from union bound}) \\ &= m \cdot \alpha \end{aligned}$$
- Since we want $m \cdot \alpha \leq \alpha'$, we can use $\alpha = \alpha' / m$. This is known as the **Bonferroni method**.
- The Bonferroni method may be too conservative. So, lesser power.

Hypothesis Testing: (Multiple testing)

- Issues with multiple testing:
 - Suppose we run m hypothesis tests on the data with null hypotheses $H_{01}, \dots, H_{0m}$
 - **Naïve method**: For every H_{0i} , it is rejected if the p-value is below the significance level α .
- **Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$
- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?
$$\begin{aligned} FWER(\alpha) &= \Pr[\text{at least one null hypothesis is falsely rejected}] \\ &\leq \sum_j \Pr[H_{0j} \text{ is falsely rejected}] \quad (\text{from union bound}) \\ &= m \cdot \alpha \end{aligned}$$
- Since we want $m \cdot \alpha \leq \alpha'$, we can use $\alpha = \alpha' / m$. This is known as the **Bonferroni method**.
- The Bonferroni method may be too conservative. So, lesser power.
- **Holm's step-down method**: Order $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_m$; $L = \min\{j: p_j > \frac{\alpha}{m-j+1}\}$; reject all null hypothesis where $p_j \leq p_L$

Hypothesis Testing: (Multiple testing)

- Issues with multiple testing:
 - Suppose we run m hypothesis tests on the data with null hypotheses $H_{01}, \dots, H_{0m}$
 - Naïve method**: For every H_{0i} , it is rejected if the p-value is below the significance level α .
- Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$
- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?
 $FWER(\alpha)$
 $= \Pr[\text{at least one null hypothesis is falsely rejected}]$
 $\leq \sum_j \Pr[H_{0j} \text{ is falsely rejected}]$
(from union bound)
 $= m \cdot \alpha$
- Since we want $m \cdot \alpha \leq \alpha'$, we can use $\alpha = \alpha'/m$. This is known as the **Bonferroni method**.
- The Bonferroni method may be too conservative. So, lesser power.
- Holm's step-down method**: Order $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_m$; $L = \min\{j: p_j > \frac{\alpha}{m-j+1}\}$; reject all null hypothesis where $p_j \leq p_L$



Hypothesis Testing: (Multiple testing)

- Issues with multiple testing:
 - Suppose we run m hypothesis tests on the data with null hypotheses $H_{01}, \dots, H_{0m}$
 - **Naïve method**: For every H_{0i} , it is rejected if the p-value is below the significance level α .
- **Family Wise Error Rate (FWER)**: $FWER(\alpha) = \Pr[V \geq 1]$
- Suppose we want FWER to be upper bounded by α' . What value of α should we use for individual hypothesis tests?
$$\begin{aligned} FWER(\alpha) &= \Pr[\text{at least one null hypothesis is falsely rejected}] \\ &\leq \sum_j \Pr[H_{0j} \text{ is falsely rejected}] && \text{(from union bound)} \\ &= m \cdot \alpha \end{aligned}$$
- Since we want $m \cdot \alpha \leq \alpha'$, we can use $\alpha = \alpha' / m$. This is known as the **Bonferroni method**.
- The Bonferroni method may be too conservative. So, lesser power.
- **Holm's step-down method**: Order $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_m$; $L = \min\{j: p_j > \frac{\alpha}{m-j+1}\}$; reject all null hypothesis where $p_j \leq p_L$.
 - Show that $\Pr[V \geq 1] \leq \alpha$.

Hypothesis Testing: (Multiple testing)

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- **Family Wise Error Rate (FWER)**: $FWER = \Pr[V \geq 1]$
- There is a tradeoff between **FWER** and the **power** (the ability to detect false null hypotheses) of multiple testing.
 - **Power**: The number of false hypotheses that we reject divided by the number of false hypotheses, i.e., $\frac{S}{m - m_0}$.
- Controlling the FWER at level α guarantees that the data analyst is very unlikely to reject any true null hypothesis, i.e., to have any **false positives**. However, this may cause failure to reject false null hypotheses (hence low power).

Hypothesis Testing: (Multiple testing)

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- **Family Wise Error Rate (FWER)**: $FWER = \Pr[V \geq 1]$
- There is a tradeoff between **FWER** and the **power** (the ability to detect false null hypotheses) of multiple testing.
 - **Power**: The number of false hypotheses that we reject divided by the number of false hypotheses, i.e., $\frac{S}{m - m_0}$.
- Controlling the FWER at level α guarantees that the data analyst is very unlikely to reject any true null hypothesis, i.e., to have any **false positives**. However, this may cause failure to reject false null hypotheses (hence low power).
- Lesson: Controlling FWER should not be the sole agenda in multiple testing.

Hypothesis Testing: (Multiple testing)

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- **Family Wise Error Rate (FWER)**: $FWER = \Pr[V \geq 1]$
- There is a tradeoff between **FWER** and the **power** (*the ability to detect false null hypotheses*) of multiple testing.
 - **Power**: The number of false hypotheses that we reject divided by the number of false hypotheses, i.e., $\frac{S}{m - m_0}$.
- **False Discovery Proportion (FDP)**: The ratio of false positives to total positives, i.e., $\frac{V}{R}$.

Hypothesis Testing: (Multiple testing)

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- **Family Wise Error Rate (FWER)**: $FWER = \Pr[V \geq 1]$
- There is a tradeoff between **FWER** and the **power** (*the ability to detect false null hypotheses*) of multiple testing.
 - **Power**: The number of false hypotheses that we reject divided by the number of false hypotheses, i.e., $\frac{S}{m - m_0}$.
- **False Discovery Proportion (FDP)**: The ratio of false positives to total positives, i.e., $\frac{V}{R}$.
- **False discovery Rate**: $FDR = E[FDP] = E[V/R]$ (*expectation over the population*)
 - Controlling FDR for m null hypothesis at $q = 0.2$: If we repeat the experiment a huge number of times, then on average, 20% of the rejected null hypotheses will be false positive.

Hypothesis Testing: (Multiple testing)

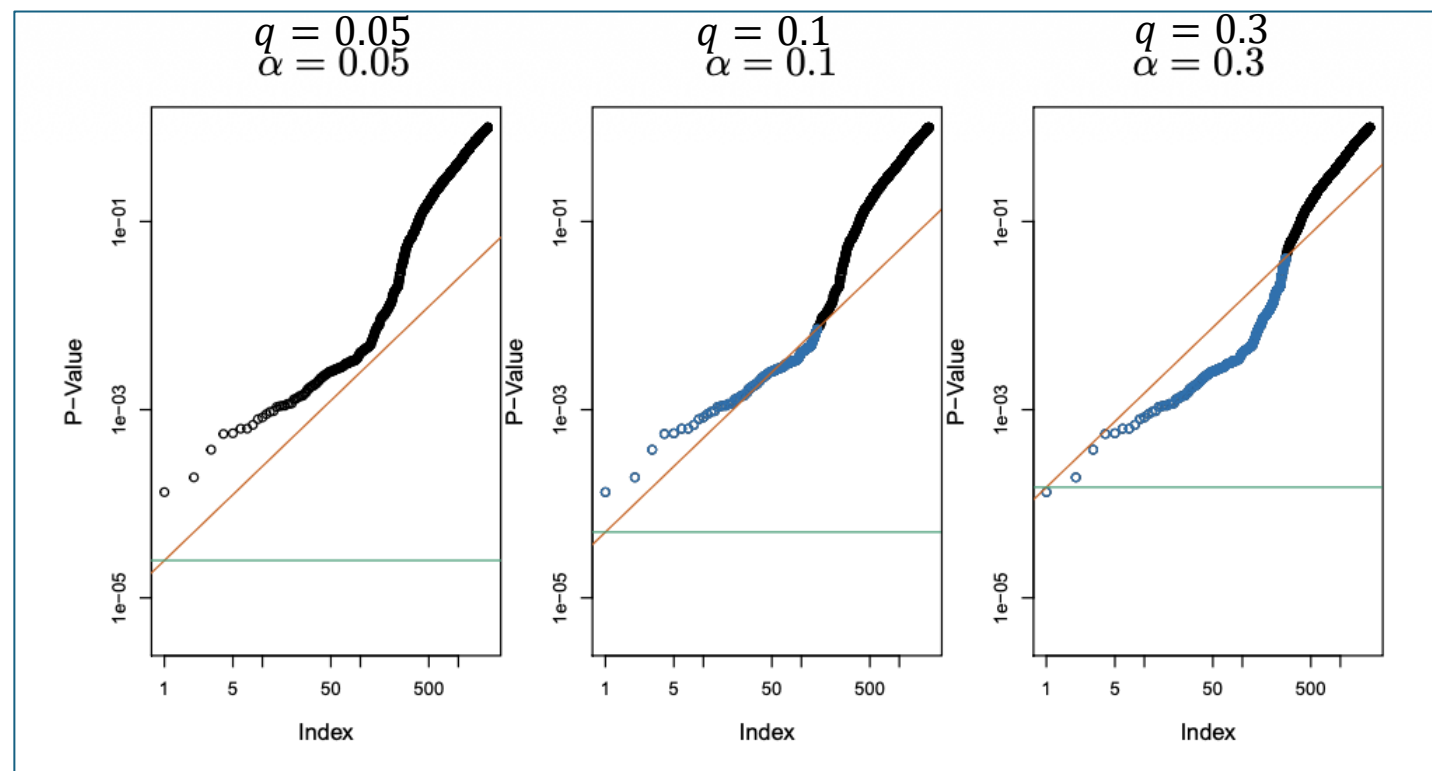
		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- **Family Wise Error Rate (FWER)**: $FWER = \Pr[V \geq 1]$
- **False discovery Rate**: $FDR = E[FDP] = E[V/R]$
- **Benjamini-Hochberg procedure**: A method to control $FDR \leq q$
 - Order $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_m$; $L = \max\{j: p_j < qj/m\}$; reject all null hypothesis where $p_j \leq p_L$.

Hypothesis Testing: (Multiple testing)

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- **Family Wise Error Rate (FWER):** $FWER = \Pr[V \geq 1]$
- **False discovery Rate:**
 $FDR = E[FDP] = E\left[\frac{V}{R}\right]$
- **Benjamini-Hochberg procedure:** A method to control $FDR \leq q$
 - Order $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_m$;
 $L = \max\{j: p_j < qj/m\}$; reject all null hypothesis where $p_j \leq p_L$.



Hypothesis Testing: (Multiple testing)

		Truth		
		Null hypothesis (H_0)	Alternate Hypothesis (H_a)	Total
Decision	Reject H_0	V	S	R
	Do not reject H_0	U	W	$m - R$
	Total	m_0	$m - m_0$	m

- **Family Wise Error Rate (FWER):** $FWER = \Pr[V \geq 1]$
- **False discovery Rate:** $FDR = E[FDP] = E\left[\frac{V}{R}\right]$
- **Benjamini-Hochberg procedure:** A method to control $FDR \leq q$
 - Order $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_m$; $L = \max\{j: p_j < qj/m\}$; reject all null hypothesis where $p_j \leq p_L$.
 - Informal result: Under mild assumptions on the p-values, it can be shown that for the above procedure $FDR \leq q$.

Parameter Estimation

Parameter Estimation

- Problem: Estimate an unknown parameter, which is a numerical value associated with the population.
 - *If the relevant data is available for the entire population, then simply calculate the parameter.*

Parameter Estimation

- Problem: Estimate an unknown parameter, which is a numerical value associated with the population.
 - *If the relevant data is available for the entire population, then simply calculate the parameter.*
 - **Issue**: Only data from a random sample is available.
 - *Reason for limited data*: The population (e.g., all people in a country) may be huge.

Parameter Estimation

- Problem: Estimate an unknown parameter, which is a numerical value associated with the population.
 - *If the relevant data is available for the entire population, then simply calculate the parameter.*
 - **Issue**: Only data from a random sample is available.
 - *Reason for limited data*: The population (e.g., all people in a country) may be huge.
 - **Solution**: *Estimate* the parameter from the random sample.
 - **Question**: *How do we quantify the error in our estimate?*

Parameter Estimation

- Problem: Estimate an unknown *parameter*, which is a *numerical* value associated with the population.
 - **Issue**: Only data from a random sample is available.
 - **Solution**: *Estimate* the parameter from the *random sample*.
 - **Question**: *How do we quantify the error in our estimate?*
- Does a single random sample provide a good estimate of the parameter?

Parameter Estimation

- Problem: Estimate an unknown *parameter*, which is a *numerical* value associated with the population.
 - Issue: Only data from a random sample is available.
 - Solution: *Estimate* the parameter from the *random sample*.
 - Question: *How do we quantify the error in our estimate?*
- Does a single random sample provide a good estimate of the parameter?
 - Typically, no.
 - Solution: Obtain multiple random samples.
 - Issue: It is typically very difficult to return to the population and obtain another sample.
 - Solution: We can try to build some *confidence* in the estimate using the single sample. Instead of a single value, we can try finding an *interval* containing the parameter.
 - Let us see the method known as “*the Bootstrap*,” which provides an interval estimate.

Parameter Estimation

- Problem: Estimate an unknown *parameter*, which is a *numerical* value associated with the population.
 - Issue: Only data from a random sample is available.
 - Solution: *Estimate* the parameter from the *random sample*.
 - Question: *How do we quantify the error in our estimate?*
- Does a single random sample provide a good estimate of the parameter?
 - Typically, no.
 - Solution: Obtain multiple random samples.
 - Issue: It is typically very difficult to return to the population and obtain another sample.
 - Solution: We can try to build some *confidence* in the estimate using the single sample. Instead of a single value, we can try finding an *interval* containing the parameter.
 - Let us see the method known as “*the Bootstrap,*” which provides an interval estimate.
 - Here is the Bootstrap method for finding a *95% Confidence Interval*:
 - Treat the original sample as if it were the population.
 - Draw from the sample, *at random with replacement, the same number of times* as the original sample size.
 - Output the interval corresponding to the “*middle 95%*” of the *resample estimates*.

Parameter Estimation

- **Problem:** Estimate an unknown *parameter*, which is a *numerical* value associated with the population.
 - **Issue:** Only data from a random sample is available.
 - **Solution:** *Estimate* the parameter from the *random sample*.
 - **Question:** *How do we quantify the error in our estimate?*
- How do you find the interval corresponding to “middle 95%” of values?
- **p^{th} percentile:** For a number p between 0 and 100, the p^{th} percentile of a collection is the smallest value in the collection that is at least as large as $p\%$ of all the values.
- Algorithm for finding the p^{th} percentile:
 - Sort the collection in increasing order.
 - Find $p\%$ of n : $(p/100) \times n$. Call that k .
 - If k is an integer, take the k^{th} element of the sorted collection.
 - If k is not an integer, round it up to the next integer, and take that element of the sorted collection.

Parameter Estimation

- **Problem:** Estimate an unknown *parameter*, which is a *numerical* value associated with the population.
 - **Issue:** Only data from a random sample is available.
 - **Solution:** *Estimate* the parameter from the *random sample*.
 - **Question:** *How do we quantify the error in our estimate?*
- How do you find the interval corresponding to “middle 95%” of values?
 - `values.percentile(97.5)` to `values.percentile(2.25)`
- **p^{th} percentile:** For a number p between 0 and 100, the p^{th} percentile of a collection is the smallest value in the collection that is at least as large as $p\%$ of all the values.
- Algorithm for finding the p^{th} percentile:
 - Sort the collection in increasing order.
 - Find $p\%$ of n : $(p/100) \times n$. Call that k .
 - If k is an integer, take the k^{th} element of the sorted collection.
 - If k is not an integer, round it up to the next integer, and take that element of the sorted collection.

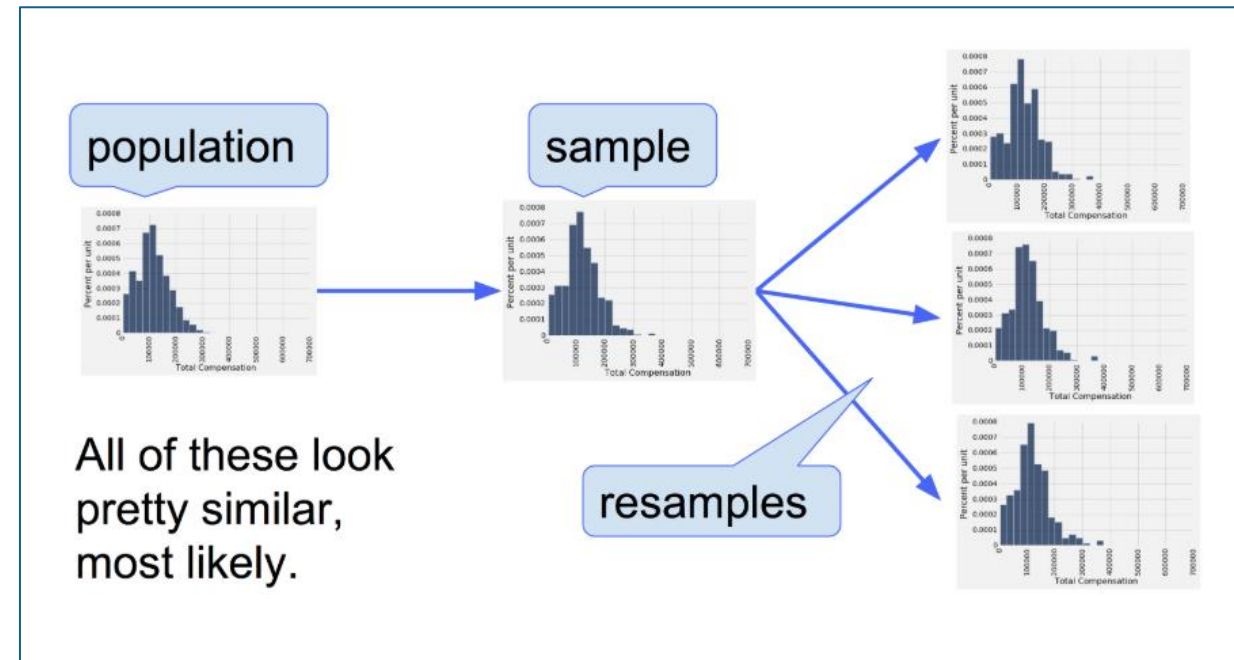
Parameter Estimation

- **Problem:** Estimate an unknown *parameter*, which is a *numerical* value associated with the population.

- **Issue:** Only data from a random sample is available.
- **Solution:** *Estimate* the parameter from the *random sample*.
- **Question:** *How do we quantify the error in our estimate?*

- Does a single random sample provide a good estimate of the parameter?

- Typically, no.
- **Solution:** Obtain multiple random samples.
- **Issue:** It is typically very difficult to return to the population and obtain another sample.
- **Solution:** We can try to build some *confidence* in the estimate using the single sample. Instead of a single value, we can try finding an *interval* containing the parameter.
- Let us see the method known as “*the Bootstrap,*” which provides an interval estimate.
- Here is the Bootstrap method for finding a *95% Confidence Interval*:
 - Treat the original sample as if it were the population.
 - Draw from the sample, *at random with replacement, the same number of times* as the original sample size.
 - Output the interval corresponding to the “*middle 95%*” of the *resample estimates*.



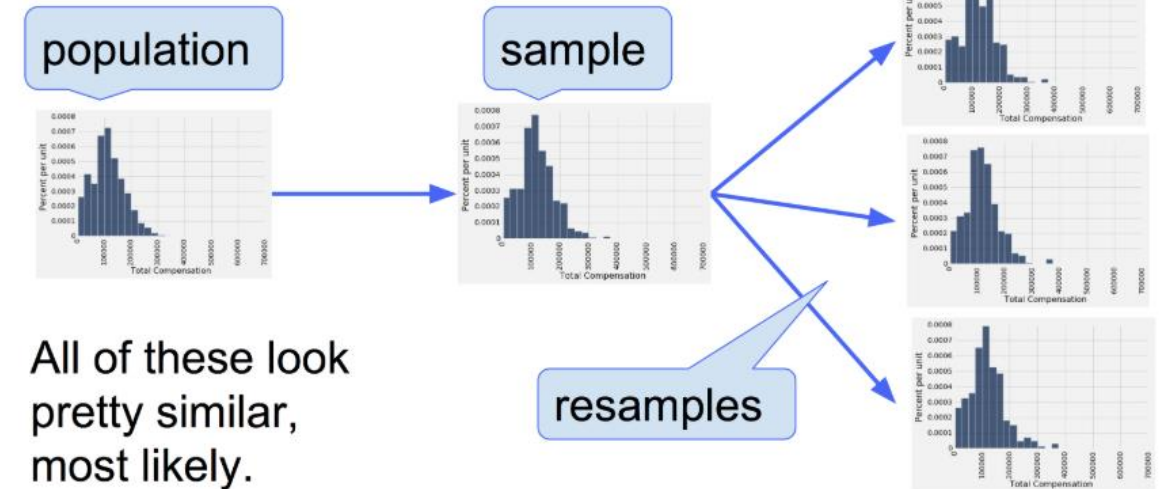
Parameter Estimation

- **Problem:** Estimate an unknown *parameter*, which is a *numerical* value associated with the population.

- **Issue:** Only data from a random sample is available.
- **Solution:** *Estimate* the parameter from the *random sample*.
- **Question:** *How do we quantify the error in our estimate?*

- Does a single random sample provide a good estimate of the parameter?

- Typically, no.
- **Solution:** Obtain multiple random samples.
- **Issue:** It is typically very difficult to return to the population and obtain another sample.
- **Solution:** We can try to build some *confidence* in the estimate using the single sample. Instead of a single value, we can try finding an *interval* containing the parameter.
- Let us see the method known as “*the Bootstrap,*” which provides an interval estimate.
- Here is the Bootstrap method for finding a *95% Confidence Interval*:
 - Treat the original sample as if it were the population.
 - Draw from the sample, *at random with replacement, the same number of times* as the original sample size.
 - Output the interval corresponding to the “*middle 95%*” of the *resample estimates*.



- **Key observation:** *If we were to repeat the experiment (i.e., take different samples), then around 95% of the interval estimates would contain the true population parameter.*

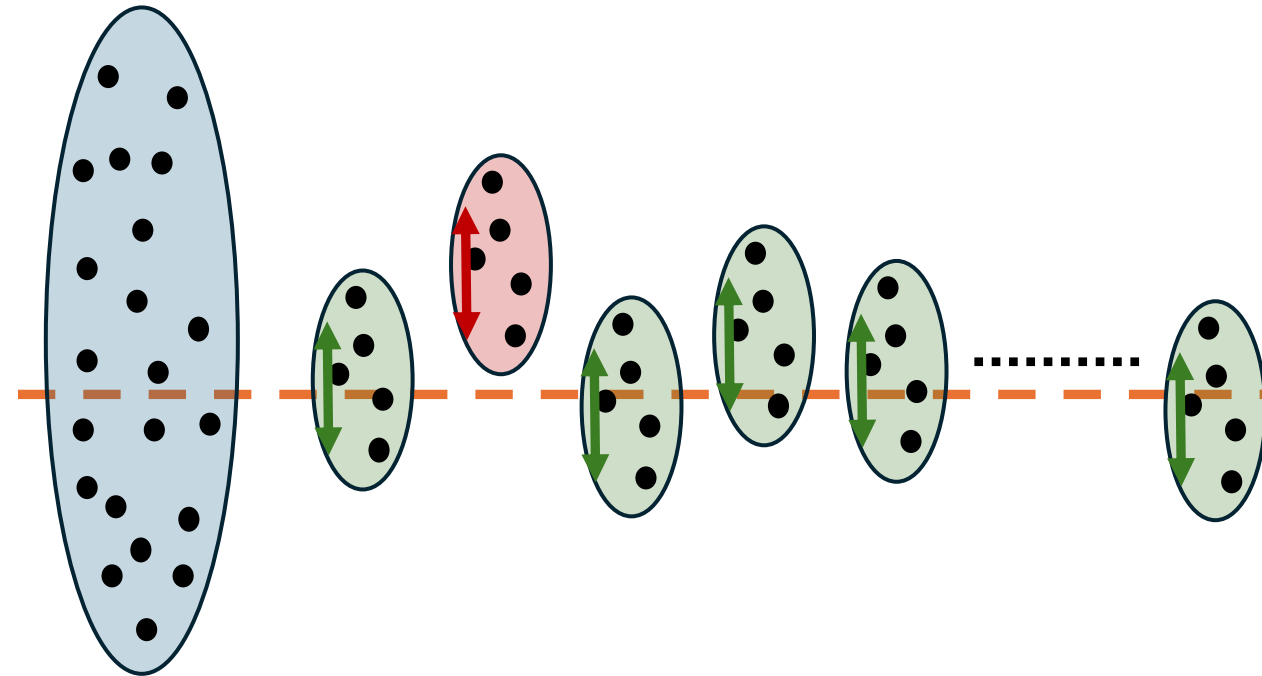
Parameter Estimation

- **Problem:** Estimate an unknown *parameter*, which is a *numerical* value associated with the population.

- **Issue:** Only data from a random sample is available.
- **Solution:** *Estimate* the parameter from the *random sample*.
- **Question:** *How do we quantify the error in our estimate?*

- Does a single random sample provide a good estimate of the parameter?

- Typically, no.
- **Solution:** Obtain multiple random samples.
- **Issue:** It is typically very difficult to return to the population and obtain another sample.
- **Solution:** We can try to build some *confidence* in the estimate using the single sample. Instead of a single value, we can try finding an *interval* containing the parameter.
- Let us see the method known as “*the Bootstrap*,” which provides an interval estimate.
- Here is the Bootstrap method for finding a *95% Confidence Interval*:
 - Treat the original sample as if it were the population.
 - Draw from the sample, *at random with replacement, the same number of times* as the original sample size.
 - Output the interval corresponding to the “*middle 95%*” of the *resample estimates*.



- **Key observation:** *If we were to repeat the experiment (i.e., take different samples), then around 95% of the interval estimates would contain the true population parameter.*

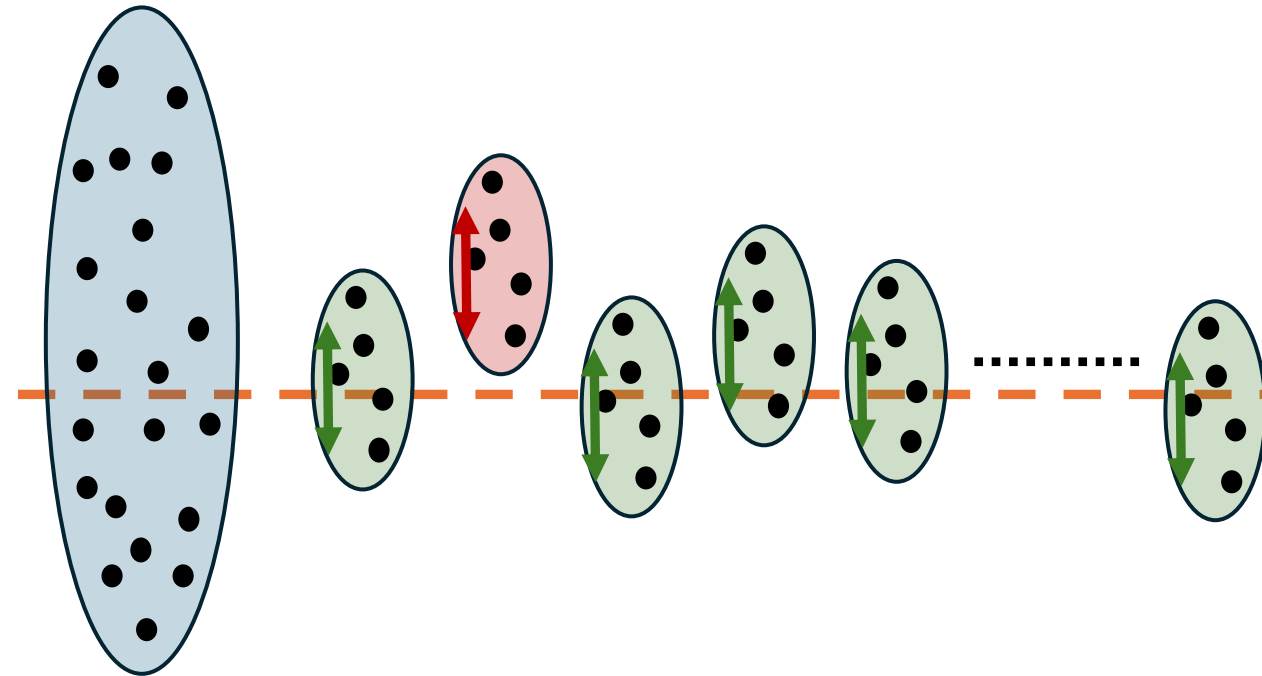
Parameter Estimation

- **Problem:** Estimate an unknown *parameter*, which is a *numerical* value associated with the population.

- **Issue:** Only data from a random sample is available.
- **Solution:** *Estimate* the parameter from the *random sample*.
- **Question:** *How do we quantify the error in our estimate?*

- Does a single random sample provide a good estimate of the parameter?

- Typically, no.
- **Solution:** Obtain multiple random samples.
- **Issue:** It is typically very difficult to return to the population and obtain another sample.
- **Solution:** We can try to build some *confidence* in the estimate using the single sample. Instead of a single value, we can try finding an *interval* containing the parameter.
- Let us see the method known as “*the Bootstrap*,” which provides an interval estimate.
- Here is the Bootstrap method for finding a *95% Confidence Interval*:
 - Treat the original sample as if it were the population.
 - Draw from the sample, *at random with replacement, the same number of times* as the original sample size.
 - Output the interval corresponding to the “*middle 95%*” of the *resample estimates*.

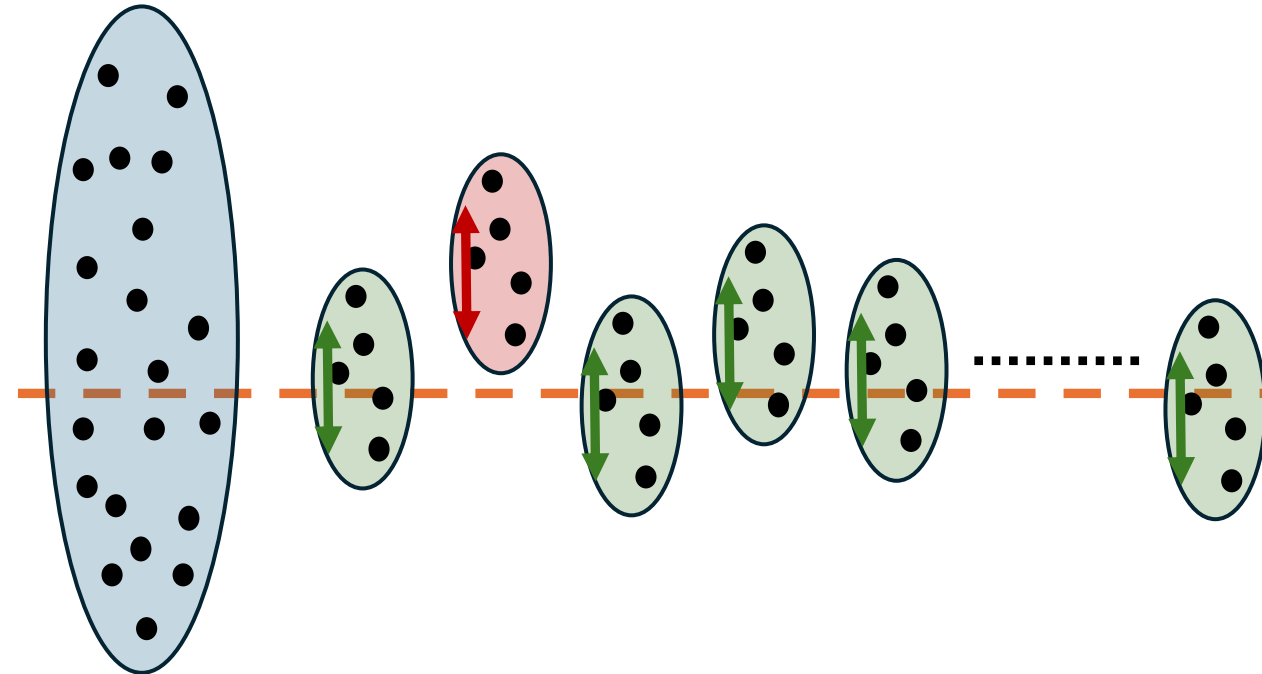


- **Common misinterpretation:** *There is a 95% chance that the true population parameter is within the interval.*

Parameter Estimation

- **Problem:** Estimate an unknown *parameter*, which is a *numerical* value associated with the population.
 - **Issue:** Only data from a random sample is available.
 - **Solution:** *Estimate* the parameter from the *random sample*.
 - **Question:** *How do we quantify the error in our estimate?*

- Does a single random sample provide a good estimate of the parameter?
 - Typically, no.
 - **Solution:** Obtain multiple random samples.
 - **Issue:** It is typically very difficult to return to the population and obtain another sample.
 - **Solution:** We can try to build some *confidence* in the estimate using the single sample. Instead of a single value, we can try finding an *interval* containing the parameter.
 - Let us see the method known as “*the Bootstrap,*” which provides an interval estimate.
 - Here is the Bootstrap method for finding a *95% Confidence Interval*:
 - Treat the original sample as if it were the population.
 - Draw from the sample, *at random with replacement, the same number of times* as the original sample size.
 - Output the interval corresponding to the “*middle 95%*” of the *resample estimates*.

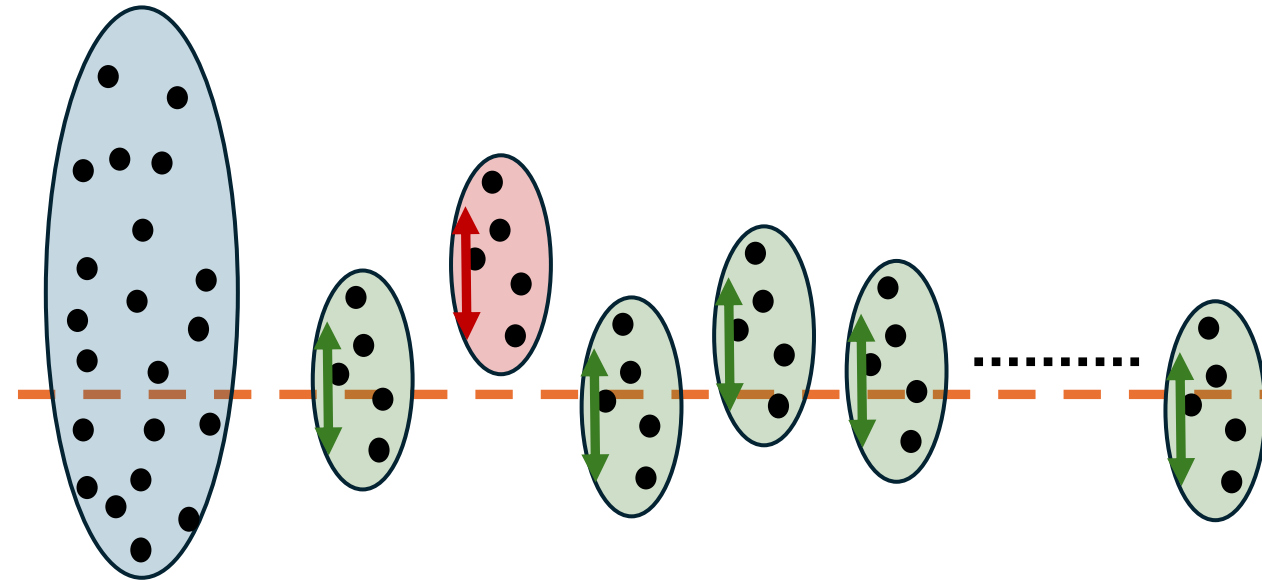


- Other comments:
 - Works well with mean/median.
 - Doesn't work for extreme parameters, e.g., max/min
 - Works well when the sample size is not too small.

Parameter Estimation

- **Problem:** Estimate an unknown *parameter*, which is a *numerical* value associated with the population.
 - **Issue:** Only data from a random sample is available.
 - **Solution:** *Estimate* the parameter from the *random sample*.
 - **Question:** *How do we quantify the error in our estimate?*

- Does a single random sample provide a good estimate of the parameter?
 - Typically, no.
 - **Solution:** Obtain multiple random samples.
 - **Issue:** It is typically very difficult to return to the population and obtain another sample.
 - **Solution:** We can try to build some *confidence* in the estimate using the single sample. Instead of a single value, we can try finding an *interval* containing the parameter.
 - Let us see the method known as “*the Bootstrap*,” which provides an interval estimate.
 - Here is the Bootstrap method for finding a *95% Confidence Interval*:
 - Treat the original sample as if it were the population.
 - Draw from the sample, *at random with replacement, the same number of times* as the original sample size.
 - Output the interval corresponding to the “*middle 95%*” of the *resample estimates*.



- Interesting duality between p-value and CI:
 - *If you are testing whether or not the population mean is a particular value x , and you use the 5% cutoff for the p-value, then you will reject the null hypothesis if x is not in your 95% confidence interval for the mean.*

End