

# Economics 403A

Assignment Project Exam Help

<https://powcoder.com>

Statistical Inference  
Add WeChat powcoder

Dr. Randall R. Rojas

# Today's Class

- Introduction to Inference
- Inference **Assignment Project Exam Help**  
**https://powcoder.com**
- Statistical Models  
**Add WeChat powcoder**
- Data Collection
- Basic Inferences

# Introduction to Inference

## Statistical Inference

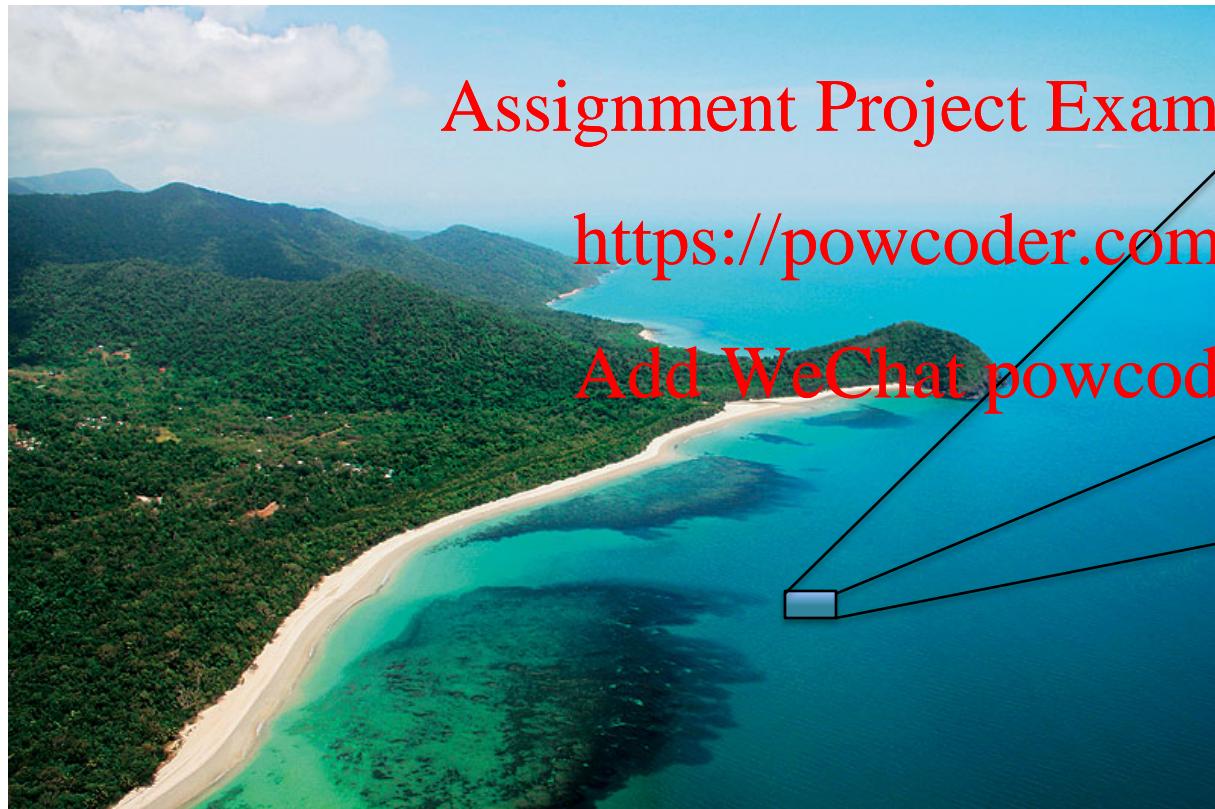
- **Def:** Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution.
  - Inference requires assumptions about the data, where different sets of assumptions represent the different models we would consider
- Types of Models:
  - a) Fully Parametric: A PDF is assumed for the DGP
  - b) Non-Parametric: Very limited assumptions made about the DGP
  - c) Semi-Parametric: These models are between a) and c)

# Introduction to Inference

Assignment Project Exam Help

<https://powcoder.com>

Add WeChat powcoder



# Inference using Probability (Fully Parametric)

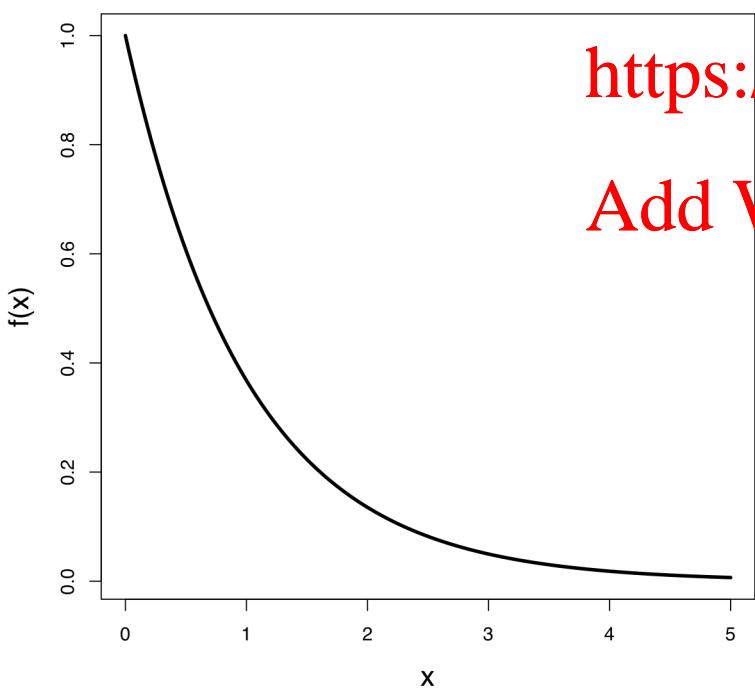
- Example (1)\*: Should I purchase an extended warranty for my new washing machine?  
Assume you want to keep it for 5–6 years.
  - Washing machine costs \$650 and the extended warranty is \$125
  - Manufacturer's warranty is good for 3 years
  - Extended warranty covers an added 5 years
- Q: How can we make an informed decision given the amount of uncertainty we face?

\*Note: See problem 5.2.1 from Evans & Rosenthal

# Inference using Probability (Fully Parametric)

- **Solution:** Assume the lifelength  $X$  in years of the washing machine follows an  $\exp(\lambda = 1)$  distribution.

Assignment Project Exam Help



- According to our model, the expected lifelength for a new machine would be  $E[X] = 1$  yr.
- The smallest interval containing 95% of the probability for  $X$  is  $(0, c)$ , where  $c$  satisfies  $0.95 = \int_0^c e^{-x} dx = 1 - e^{-c} \rightarrow c \approx 3$  is equal to  $[0, 3]$ .
- How likely is it that it will still work after 5 yrs?

$$P(X \geq 5) = \int_5^\infty e^{-x} dx = 0.0067$$

Better get that warranty! 😬

# Inference using Probability (Fully Parametric)

- Example (1): Suppose you didn't get the extended warranty, and a year has past but so far its working fine. Would your previous results change?  
Assignment Project Exam Help

<https://powcoder.com>

$$E(X | X > 1) = \int_1^\infty x e^{-x} dx = -x e^{-x} \Big|_1^\infty + \int_1^\infty e^{-(x-1)} dx = 2.$$

$$0.95 = \int_1^c e^{-(x-1)} dx = e(e^{-1} - e^{-c}) \rightarrow c = -\ln(e^{-1} - 0.95e^{-1}) = 3.9957$$

$$P(X > 5 | X > 1) = \int_5^\infty e^{-(x-1)} dx = e^{-4} = 0.0183$$

# Statistical Models

- **Example:** Suppose that you work for an insurance company and historically they found that the number of claims they receive can be described by a Poisson distribution but the parameter value (i.e.,  $\theta$ ) is unknown.  
**Assignment Project Exam Help**  
**<https://powcoder.com>**  
– Recall that:  $P(X = x|\theta) = \frac{\theta^x e^{-\theta}}{x!}$   
**Add WeChat powcoder**
- They assign you the task of figuring out the probability that 6 claims will be made tomorrow given that today 4 were received.
- **Q:** How should you proceed?

# Statistical Models

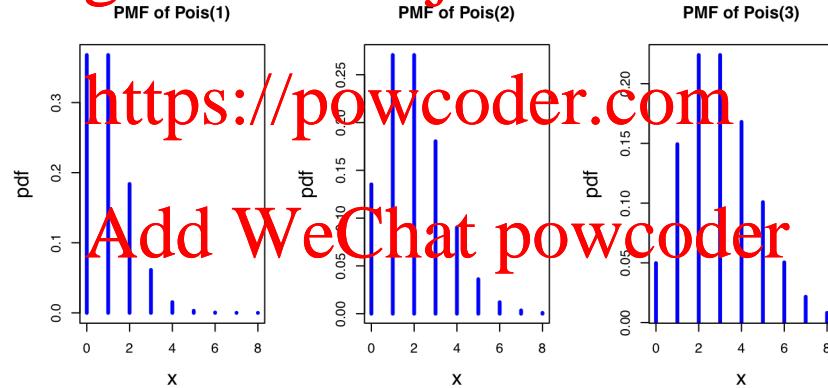
- Solution:
  1. Try a ‘brut force approach’ by considering all the values  $\theta$  can take, e.g.,  $\theta = 1, \theta = 2, \dots$ , and for each one, computing the respective PMF.  
<https://powcoder.com>
  2. From the family of PMFs, identify the one(s) most ‘consistent’ with the data.
  3. Based on your PMF(s) from Step 2, compute the desired probability  $P(X = 6|\theta)$ .

# Statistical Models

- Step 1:

Try a ‘brut force approach’ by considering all the values  $\theta$  can take, e.g.,  $\theta=1, \theta=2, \dots$ , and for each one, computing the respective PMF.

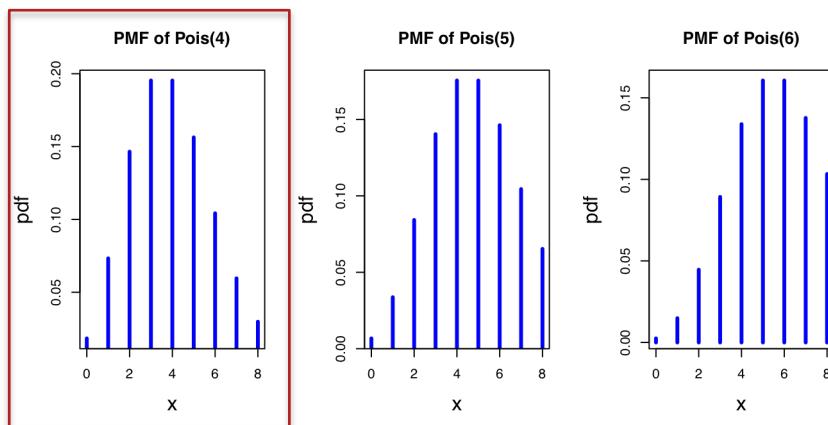
Assignment Project Exam Help



https://powcoder.com  
Add WeChat powcoder

- Step 2:

Data: { $X = 4$ }  
 $\rightarrow P(X=4 | \theta=4) \approx 0.2$



# Statistical Models

- Step 3:

Based on your PMF(s) from Step 2, compute the  
desired probability  $P(X = 6|\theta)$ .

<https://powcoder.com>

From the figures we can try finding  $P(X=6)$  using the parameter value  $\theta=4$ , i.e.,  $P(X=6|\theta=4)$ :

$$\rightarrow P(X=6 | \theta=4) = 0.10.$$

# Statistical Models

## Definitions

- Parameter Space:  $\Omega = \{\theta_1, \theta_2, \dots, \theta_n\}$   
Is the set of all possible values of the parameter  $\theta$ .  
**Assignment Project Exam Help**
- Data: Observations ( $= x$ ) obtained from a random mechanism ( $= P$  = ‘probability measure’) assumed to have generated them.  
**Add WeChat powcoder**
- Statistical Model: Represents our choice of probability measure from the many plausible ones, such that given  $\theta \in \Omega$ ,  $P_\theta$  is the true probability measure.

**Goal:** We observe the data but not  $P$ , yet we want to make inferences about  $P$

# Statistical Models

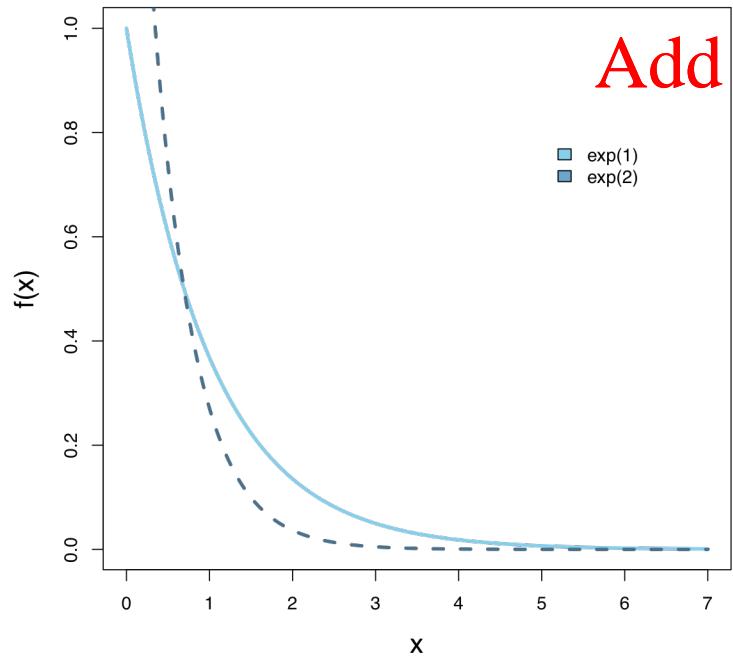
- In practice, we often have more than just one observation, and the distribution may depend on more than one parameter.

[Assignment Project Exam Help  
https://powcoder.com](https://powcoder.com)

- The previous formalism can be easily generalized to a vector of parameters  $\vec{\theta}$  and set of observations  $(x_1, \dots, x_n)$ .

# Statistical Models

- Example (see 5.3.2): Suppose there are two manufacturing plants for machines. It is known that machines built by the first plant have lifelengths distributed  $\text{Exponential}(1)$ , while machines manufactured by the second plant have lifelengths distributed  $\text{Exponential}(2)$ .
- You have purchased five of these machines knowing that all five came from the same plant, but you do not know which plant.



Add WeChat powcoder

Statistical Model:  $\{P_1, P_2\}$   
Parameter Space:  $\Omega = \{1,2\}$

If we observe  $(x_1, \dots, x_5) =$

$$\left. \begin{array}{l} (5.0, 3.5, 3.3, 4.1, 2.8) \\ \quad \rightarrow \theta = 2 \\ (2.0, 2.5, 3.0, 3.1, 1.8) \\ \quad \rightarrow \theta = 1 \end{array} \right\}$$

# Data Collection

- If the sample is large enough and representative of the degree of variation of the population, you do not need to use the entire population  
**Assignment Project Exam Help**  
<https://powcoder.com>
- There are 4\* common forms of ‘Probability Sampling’ from a population  
**Add: WeChat powcoder**
  - a) Simple Random Sampling
  - b) Stratified Sampling (R package = splitstackshape)
  - c) Cluster & Multistage Sampling (R library = survey)
  - d) Systematic Sampling

\*Note: See De Veaux, Velleman & Bock, Intro Stats (Ch 12) for more details

# Data Collection

## (A) Simple Random Sampling (SRS)

- We draw samples because we can't work with the entire population.  
**Assignment Project Exam Help**
  - We need to be sure that the statistics we compute from the sample reflect the corresponding parameters accurately.
  - A sample that **Add WeChat powcoder** is said to be **representative**.
- Example: The IRS may audit a random sample of 500 tax returns from a small county with 20,000 people.
  - This is financially and practically more feasible to do than to inspect all 20,000 returns

# Data Collection

## (B) Stratified Sampling

- Simple random sampling is not the only fair way to sample. **Assignment Project Exam Help**
- More complicated designs may save time or money or help avoid sampling problems.
- All statistical sampling designs have in common the idea that chance, rather than human choice, is used to select the sample.
- Designs used to sample from large populations are often more complicated than simple random samples.

# Data Collection

## (B) Stratified Sampling

- Sometimes the population is first sliced into homogeneous groups, called strata, before the sample is selected.
- Then simple random sampling is used within each stratum before the results are combined.
- This common sampling design is called stratified random sampling.
- Example: Suppose Boeing wants to determine the level of compliance with their no-drugs consumption policy. Would an SRS of 50 employees out of 2000 be appropriate?

→ No! Some departments are much larger than others, and therefore more likely to be included in the sample.  
Instead use Stratified Sampling!

# Data Collection

## (C) Cluster and Multistage Sampling

- Sometimes stratifying is not practical and simple random sampling is difficult (e.g. too expensive to include everyone).  
**Assignment Project Exam Help**  
**https://powcoder.com**
- Splitting the population into similar parts or **clusters** can make sampling more practical.
  - Then we could select one or a few clusters at random and perform a census within each of them.
  - This sampling design is called **cluster sampling**.
  - If each cluster fairly represents the full population, cluster sampling will give us an unbiased sample.
- **Example:** In the US, are auto shops charging women more than men for the same car repairs?

# Data Collection

## (C) Cluster and Multistage Sampling

- Sometimes we use a variety of sampling methods together.
- Sampling schemes that combine several methods are called **multistage samples**.
- Most surveys conducted by professional polling organizations use some combination of stratified and cluster sampling as well as simple random sampling.

# Data Collection

## (D) Systematic Sampling

- Sometimes we draw a sample by selecting individuals systematically.  
**Assignment Project Exam Help**
  - For example, you might survey every 10th person on an alphabetical list of students.  
<https://powcoder.com>
- To make it random, you must still start the systematic selection from a randomly selected individual.
- When there is no reason to believe that the order of the list could be associated in any way with the responses sought, **systematic sampling** can give a representative sample.

# Basic Inferences

## Descriptive Statistics

- Def:  $p$ th quantile (or 100  $p$ th percentile) =  $\pi_p$   
$$p = \int_{-\infty}^{\pi_p} f(x)dx = F(\pi_p)$$
where  $F$  = CDF and  $f(x)$  = PDF.  
Assignment Project Exam Help  
<https://powcoder.com>  
Add WeChat powcoder
- Five Number Summary
  - Consists of the min,  $Q_1$ ,  $Q_2$  (median), IQR,  $Q_3$  and max, where  $Q_1 = \pi_{0.25}$ , median =  $\pi_{0.50}$ ,  $Q_3 = \pi_{0.75}$ , and IQR =  $Q_3 - Q_1$

# Basic Inferences

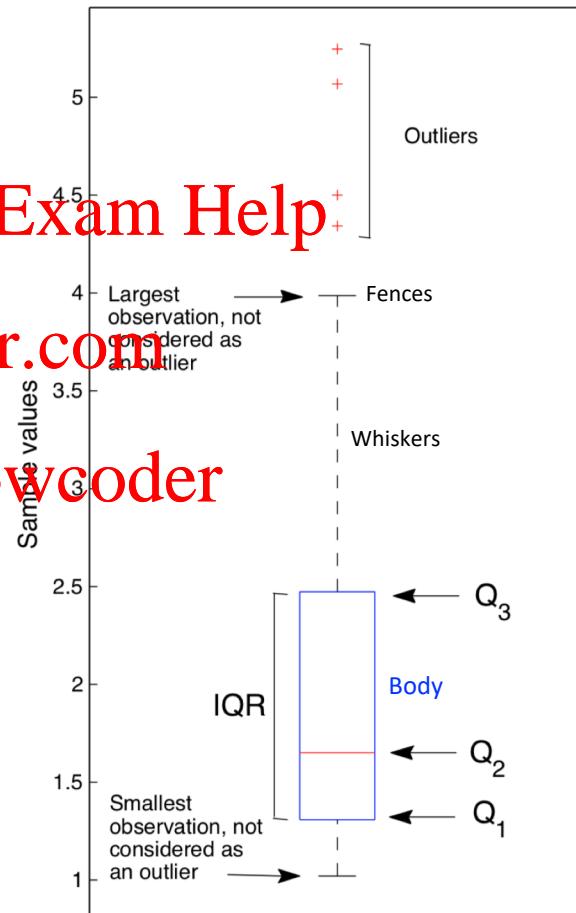
## Descriptive Statistics

- Example: Given  $f(x) = 2e^{-2x}$ ,  $x > 0$ , find  $\pi_{0.5}$ .  
[Assignment Project Exam Help](https://powcoder.com)  
<https://powcoder.com>
- Solve for  $\pi_p$  from:  $0.5 = \int_0^{\pi_0.5} 2e^{-2x} dx = F(\pi_{0.5})$   
 $\rightarrow \pi_{0.5} = \frac{1}{2} \ln(2)$

# Basic Inferences

## Descriptive Statistics

- Boxplot: Graphical representation of the number summary.
- Useful for comparing groups
- It consists of:
  - Body:  $Q_1$ ,  $Q_2$ ,  $Q_3$
  - Fences: Lower =  $Q_1 - 1.5 \text{IQR}$
  - Upper =  $Q_3 + 1.5 \text{IQR}$
  - Whiskers: dashed lines



# Basic Inferences

## Descriptive Statistics

- Histogram: Graphical representation of a quantitative variable's density distribution.
- Q: How many bins should be used?
- A: For a dataset with  $n$  observations, a nearly optimal number is  $k = 1 + \log_2(n)$

