Cross-Institutional Undergraduate Research Experience

Initiation of A Potential Research

Kedai Cheng

University of North Carolina Asheville

Outline

1 **How I Initiated My First Research**
- Tolerance
- Pointwise Tolerance Intervals for Autoregressive Models

2 **Ongoing Projects**

3 **Let's Get Started**
- Motivations
- Central Limit Theorem
- Confidence Intervals
- Hypothesis Tests

Introduction

Question:

- Do 99% of brakes manufactured by a brand have good quality?
- Is a particular developed vaccine safe and effective to 95% of the population?
- What is the price of stock X will be tomorrow? At which price will you load or unload shares? How much confidence do you hold?

Proposed methodology:

- Pointwise Tolerance Intervals for Autoregressive ($AR(p)$) Models

Motivations

- Tolerance intervals are available for numerous settings. However, approaches for autoregressive models are far more limited.
- We want to establish tolerance intervals for general $AR(p)$ models, which may also have a mean or trend component presented.
- Real application: length of waitinglists are a symbol of efficiency of hospital services. Understanding the uncertainty of forecasting growth/ decline of waitinglist could help hospital managers with capacity planning.

AR(p) Models

The autoregressive model of order p, or AR(p) model, with a deterministic trend is:

$$x_t = \mu_t + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \cdots + \varphi_p x_{t-p} + \varepsilon_t \quad (1)$$

where p is number of lags, $\varphi_i' s$ are model parameters, ε_t is a random noise, and $\mu_t = \beta_0 + \sum_{k=1}^{q-1} \beta_k g_k(t)$ is a linear trend.

Tolerance Interval

- A (P, γ) tolerance interval for the sampled population is:

$$Pr_{\bar{X},S}\{Pr_X(\bar{X} - k_{P,\alpha;f}S \leq X \leq \bar{X} + k_{P,\alpha;f}S) \geq P\} = \gamma \tag{2}$$

  Nominally, $\gamma \times 100\%$ of these tolerance intervals will capture *at least* $P \times 100\%$ of the sample population.

- If $X \sim N(\mu, \sigma^2)$, we have

$$Pr_{\bar{X},S}\left[Pr_X\left(\frac{\bar{X} - \mu - kS}{\sigma} \leq \frac{X - \mu}{\sigma} \leq \frac{\bar{X} - \mu + kS}{\sigma}\right) \geq P\right] = \gamma \tag{3}$$

Process Variance

- Amin and Li (2002) proposed usual sample variance estimate as $S^2 = \frac{1}{T-1}\sum_{t=1}^{T}(x_t - \bar{x})$ [Amin and Li(2002)]
  However, the independence assumption is violated in time series model.
- We propose using the estimator of the variance of the Gaussian error terms to be

$$S^2 = \frac{1}{T-p-q}\sum_{t=p+1}^{T}\hat{\varepsilon}_t^2 \tag{4}$$

where $\hat{\varepsilon}_t = x_t - \hat{\beta}^T\mathbf{z}_t - \sum_{i=1}^{p}\hat{\varphi}(x_{t-1} - \hat{\beta}^T\mathbf{z}_{t-i})$, and $\mathbf{z}_t$ is a $q-$dimensional vector $(1, t, g_1(t), \cdots, g_{q-2}(t))$.

Process Variance

- Without trend to be estimated, the process variance is

$$\gamma_p = \frac{\sigma_\varepsilon^2}{1 - \sum_{i=1}^{p} \rho_i \varphi_i} \tag{5}$$

- With trend to be estimated, the process variance [Cryer and Chan(2008)] is

$$\zeta_{p,t^*}^2 = \gamma_p (1 + \mathbf{z}_{t^*}^T (\mathbf{Z}_T^T \mathbf{Z}_T)^{-1} \mathbf{z}_{t^*}) \tag{6}$$

where $\mathbf{Z}_T$ is a $T \times q$ matrix with t^{th} row $\mathbf{z}_t^T$.

Adjusted k-factor

- The conventional k-factor assumes the data are *i.i.d* normal. Because of the intractability with deriving a k-factor under autoregressive settings, the k-factor proposed in equation (2) is proposed as a practical surrogate.
- However, for an AR(p) model with a deterministic linear trend, the degree of freedom need to be reflected through the adjusted k-factor. Therefore, we propose the adjusted k-factor to be $k^*_{P,\alpha} = k_{P,\alpha;T-p-q}$.
- Our proposed tolerance interval is

$$\hat{x}_t \pm k_{P,\alpha;T-p-q}\hat{\zeta}_{p,t^*} \tag{7}$$

Adjustment of α

- For $i = 1, \cdots, M$, simulate a time series of length T;
- Estimate parameters $(\hat{\beta}_i, \hat{\varphi}_i)$;
- Simulate $b = 1, \cdots, B$ bootstrap samples from $(\hat{\beta}_i, \hat{\varphi}_i)$, call them $\{x^*_{ij1}\}_{j=1}^{T}, \cdots, \{x^*_{ijB}\}_{j=1}^{T}$;
- For each sample, construct $(P, 1-\alpha^*),\ k = 1, \cdots, K$, tolerance intervals for a set of candidate α^*;
- Calculate the coverage probability for each tolerance interval, which we call $\mathcal{C}(P, 1-\alpha^*)$, find α^*_i such that

$$\alpha^*_i = argmin_{\alpha^*_k} |\mathcal{C}(P, 1-\alpha^*_k) - (1-\alpha)|$$

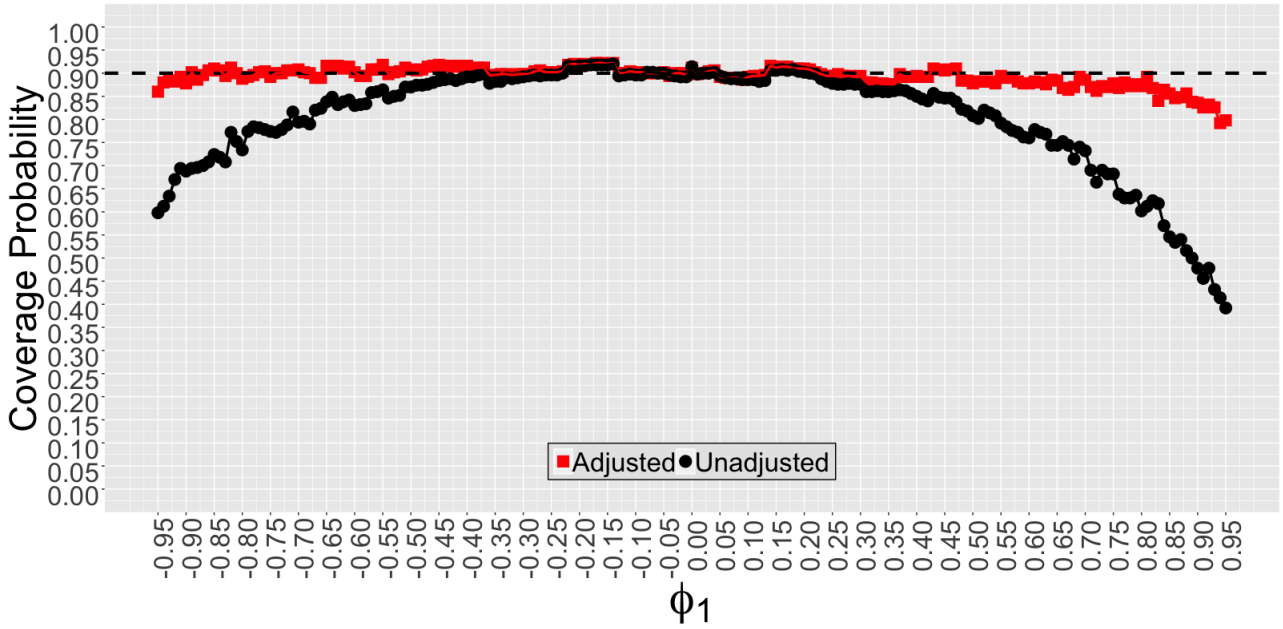
- Construct the $(P, 1-\alpha^*_i)$ tolerance limits for $\{x_{ij}\}_{j=1}^{T}$.

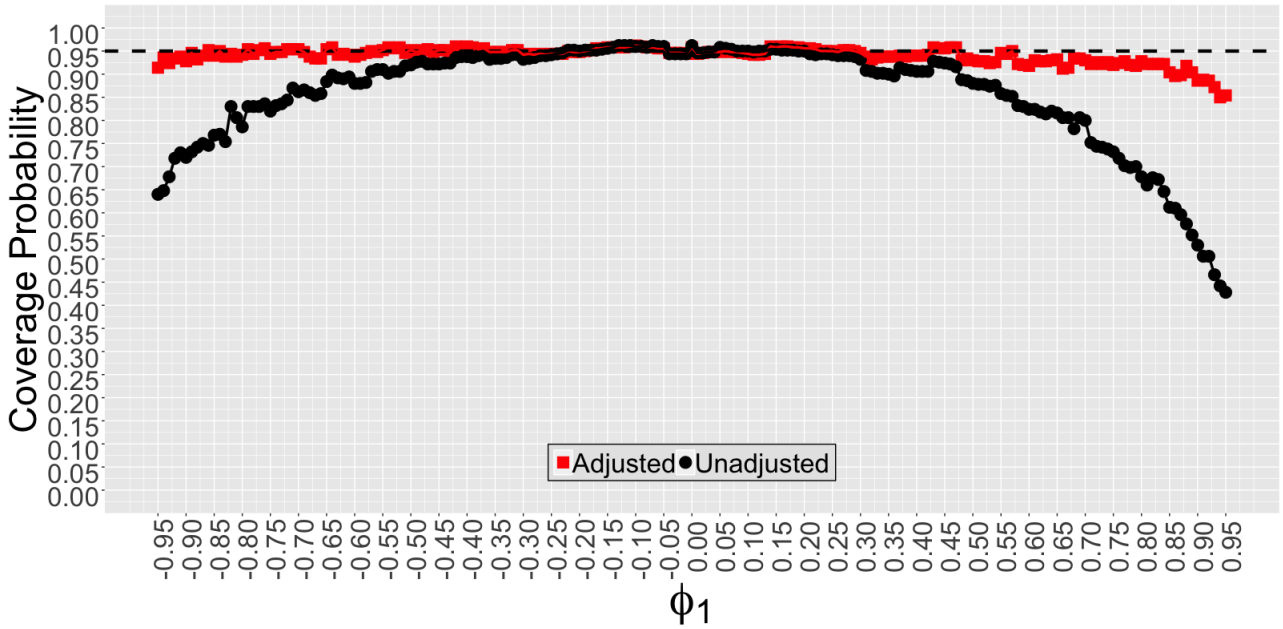
Coverage Study for $AR(1)$ Model

We conduct coverage studies for the following settings:

- $\varphi_1 = \{\pm 0.95, \pm 0.94, \cdots, \pm 0.01, 0\}$
- $(P, \gamma) = \{(0.90, 0.90),\ (0.95, 0.95)\}$
- $T = \{25, 50, 100, 500\}$
- $\mu_t = \beta_0 = 3$, and $\varepsilon_t \sim N(0, 1)$

Coverage Study for $AR(1)$ Model



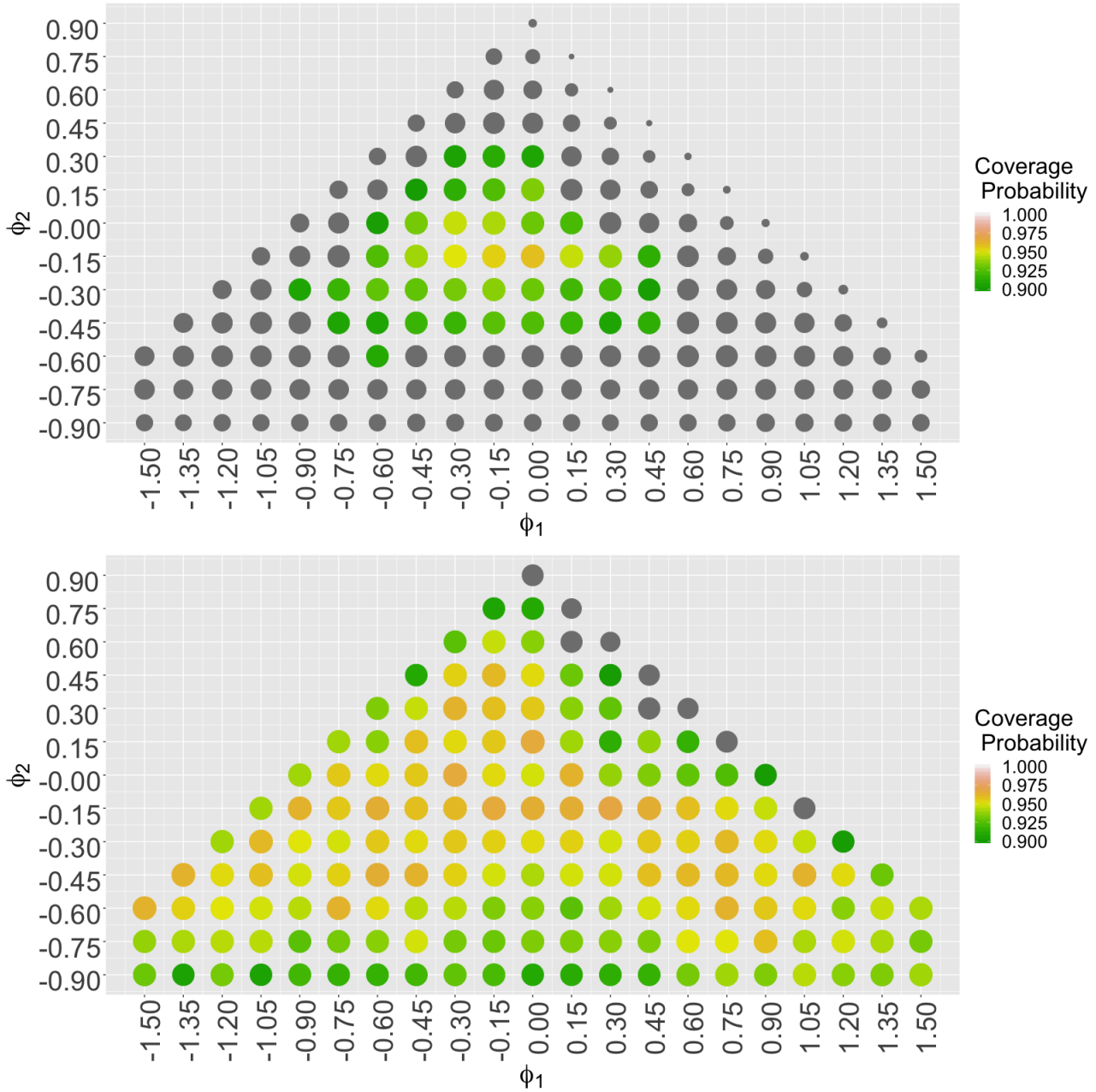


Coverage Study for $AR(2)$ Model

We conduct coverage studies for the following settings:

- $\varphi_1 = \{\pm 1.50, \pm 1.35, \cdots, \pm 0.15, 0\}$
- $\varphi_2 = \{\pm 0.90, \pm 0.75, \cdots, \pm 0.15, 0\}$
- $(P, \gamma) = (0.95, 0.95)$
- $T = \{25, 50, 100, 500\}$
- $\beta_0 = 1, \beta_1 = 3$ and $\varepsilon_t \sim N(0, 1)$

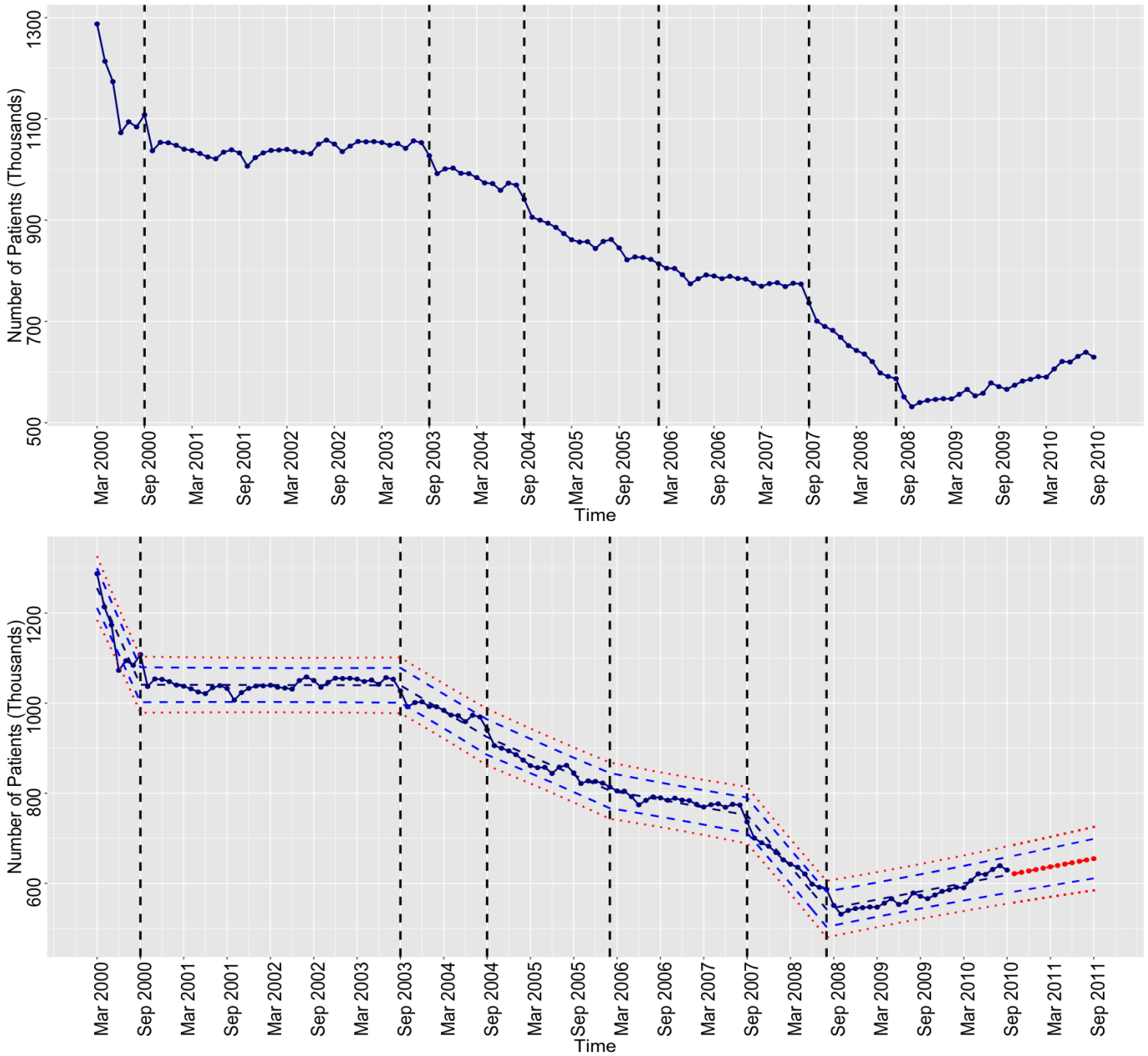
Coverage Study for $AR(2)$ Model



Application: UK Hospital Waitinglist [hos(2013)]

- Monthly number of patients on provider-based waiting lists to be admitted to NHS hospitals in England.
- T=127. Monthly data from March 2000 to September 2010.

Application: UK Hospital Waitinglist



Application: UK Hospital Waitinglist

Date	Forecast	Unadjusted TI		Adjusted TI	
		Lower Limit	Upper Limit	Lower Limit	Upper Limit
October 2010	621,523	581,546	661,500	557,453	685,593
November 2010	624,555	584,295	664,815	560,032	689,079
December 2010	627,588	587,028	668,147	562,584	692,592
January 2011	630,620	589,745	671,495	565,110	696,130
February 2011	633,652	592,446	674,859	567,611	699,693
March 2011	636,685	595,131	678,238	570,088	703,281
April 2011	639,717	597,802	681,632	572,541	706,893
May 2011	642,749	600,458	685,040	574,970	710,528
June 2011	645,782	603,100	688,463	577,377	714,186
July 2011	648,814	605,728	691,900	579,761	717,866
August 2011	651,846	608,343	695,349	582,124	721,568
September 2011	654,879	610,945	698,812	584,467	725,290

Summary

- Develop statistical tolerance intervals for general $AR(p)$ models;
- Tolerance Intervals with presence of a linear trend;
- Bootstrap adjustment to improve the coverage performance.

Some of Ongoing Projects

- Specifying Cutoffs in Trimming and Winsorization (Math Skills)
- Handbook of Regression Methods (Coding Skills)
- `tolerance` (Coding Skills)
- *Cross-Institutional Undergraduate Research Experience Workshop* (Presentation Skills)

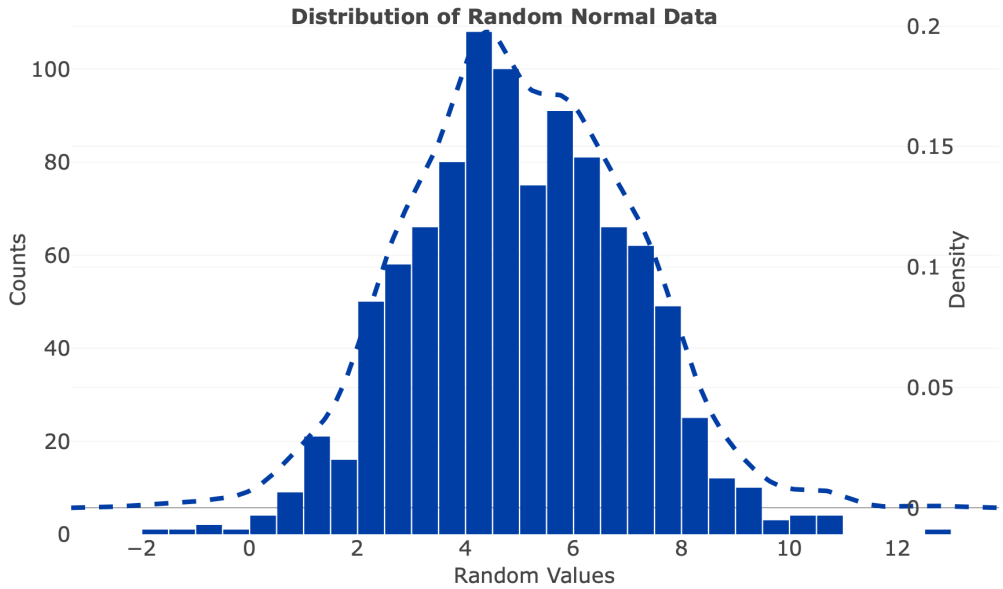
Questions

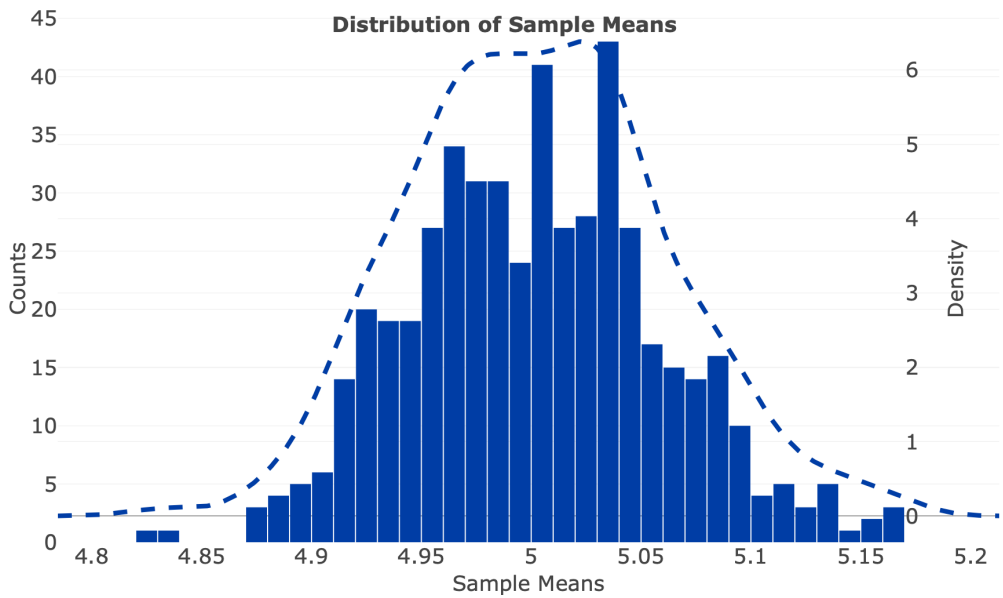
- Would you think Coke is sweeter than Pepsi?
- Would you think children is more stress resistant than adults?
- Would you think 85% people can pass their driver's license test at first time?

Think is important.

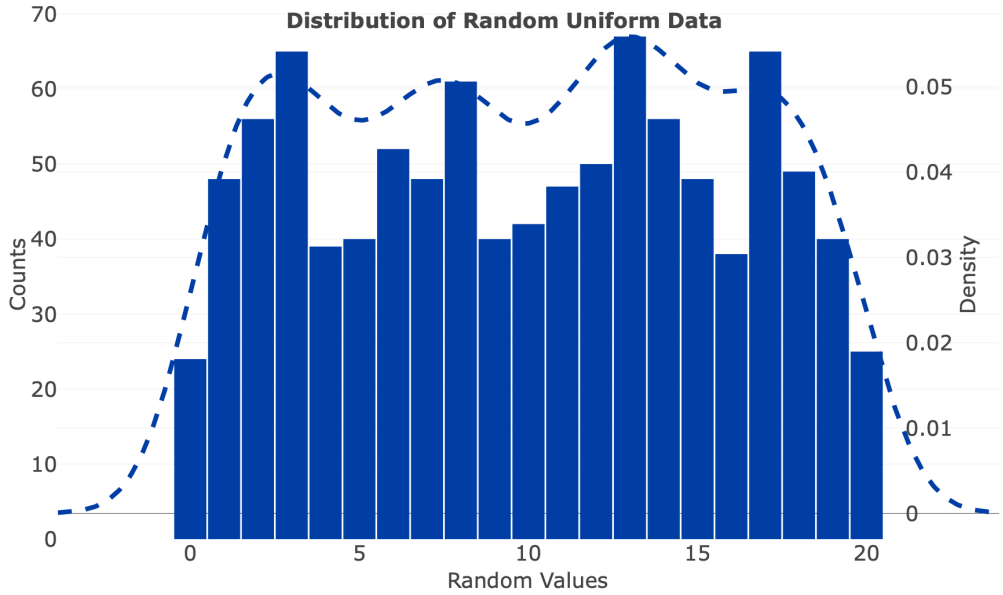
Prove is convincing.

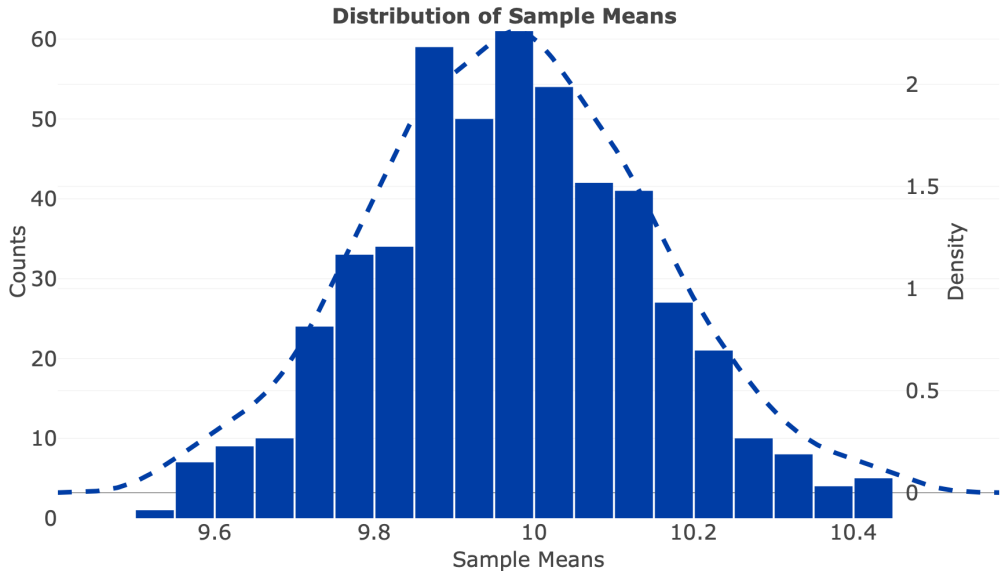
Normal Data (Code)



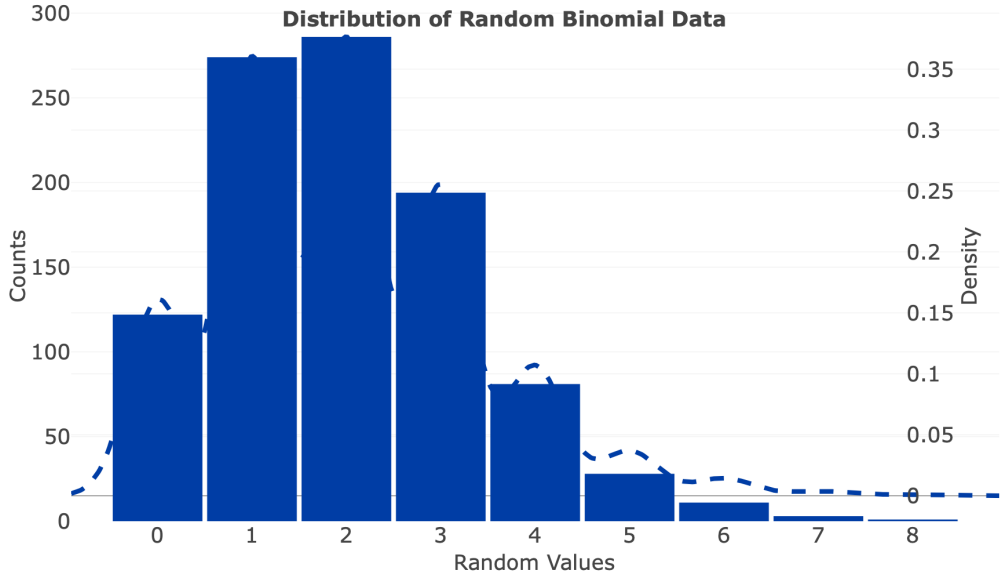


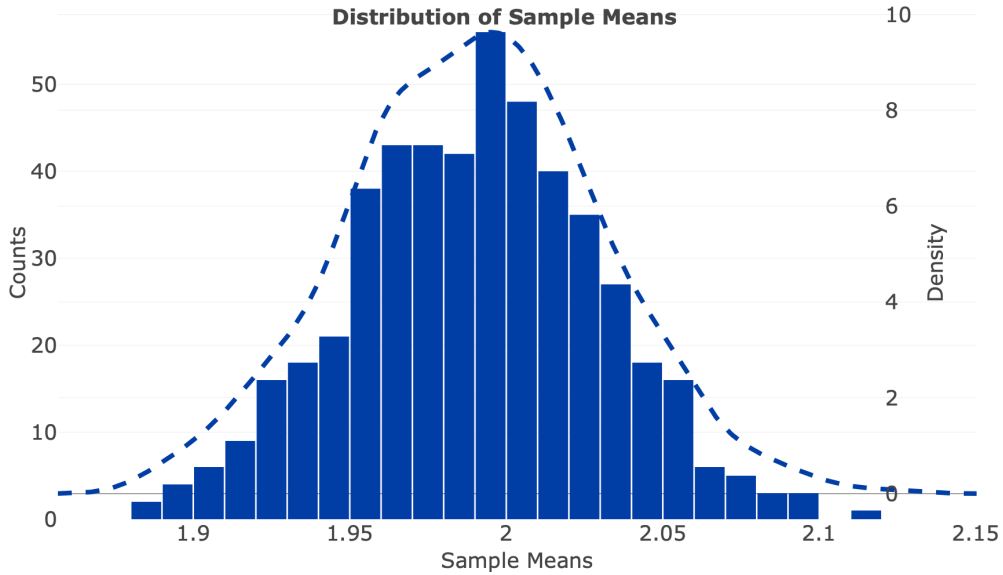
Uniform Data (Code)





Binomial Data (Code)





Central Limit Theorem

Theorem

If $\mathbf{X}$ *is a random sample from a population with mean* μ *and standard deviation* σ*, then the sample mean* $\bar{\mathbf{X}}$ *follows a normal distribution, such that*

$$\sqrt{n}\left(\frac{\bar{\mathbf{X}} - \mu}{\sigma}\right) \quad \sim \quad \mathcal{N}(0, 1)$$

Confidence Intervals

- Confidence intervals are used to capture a parameter of interest with a specified confidence level.
- Confidence intervals are constructed based on the distribution of corresponding statistics.
- Mathematically,

$$\mathbb{P}\Big(L \leq X \leq U\Big) \geq 1 - \alpha$$

where L and U denotes lower and upper confidence intervals, respectively. α is the critical level.

Confidence Intervals for Population Mean

- Two-sided Confidence Interval

$$Statistic \quad \pm \quad M.O.E$$

where $M.O.E$ stands for *margin of error*.

$$\left(\bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad , \quad \bar{x} + Z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \tag{8}$$

where Z_γ is the quantile function for a *standard normal* distribution that covers $\gamma \times 100\%$ of the population.

In practice,

$$\left(\bar{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \quad , \quad \bar{x} + t_{1-\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right) \tag{9}$$

where s is sample standard deviation, n is sample size.

Confidence Intervals for Population Mean

- One-sided Confidence Interval

$$\left(\bar{x} - Z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad , \quad +\infty\right)$$

$$\left(-\infty \quad , \quad \bar{x} + Z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$$

where Z_γ is the quantile function for a *standard normal* distribution that covers $\gamma \times 100\%$ of the population.

In practice,

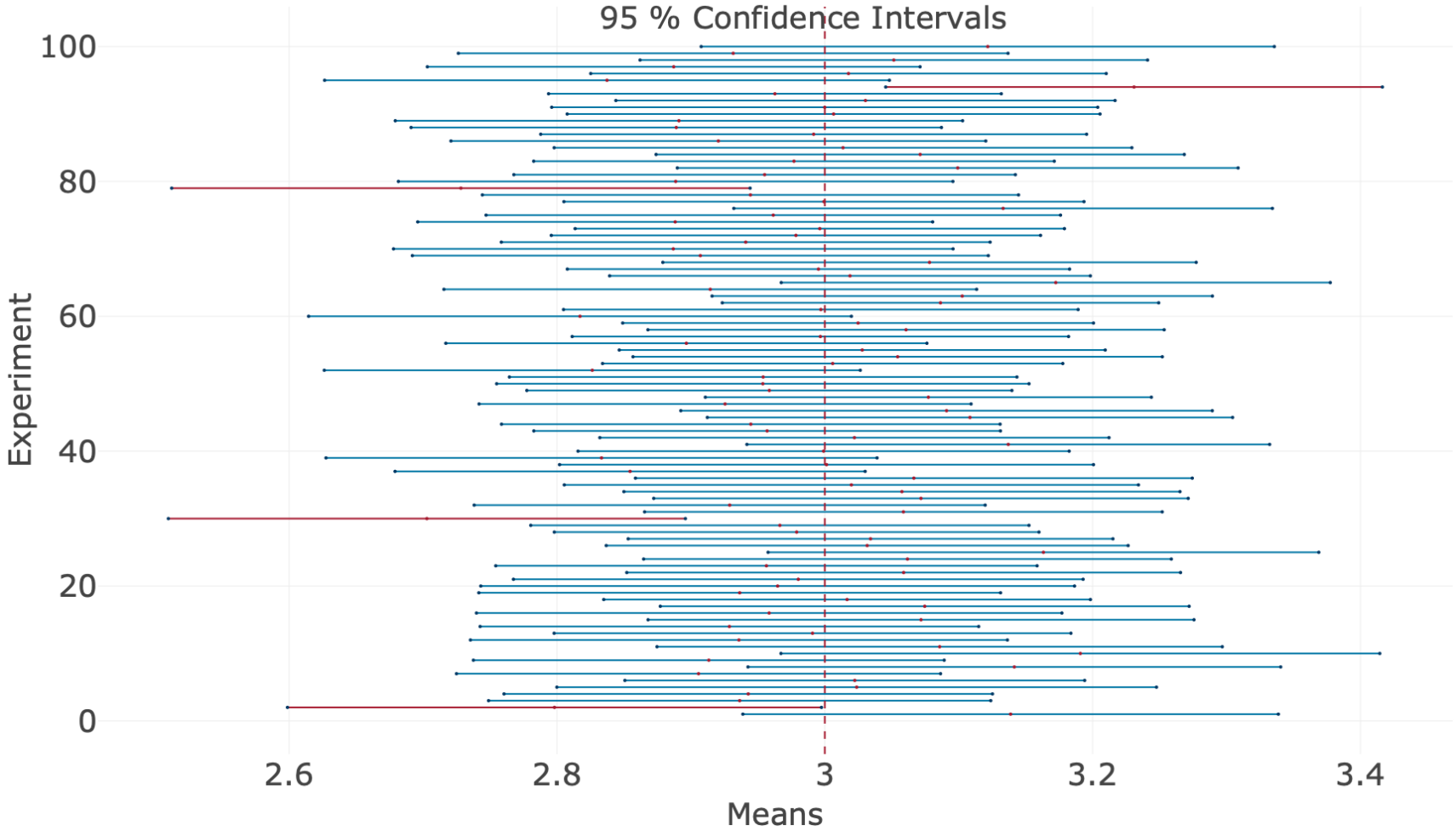
$$\left(\bar{x} - t_{1-\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} \quad , \quad +\infty\right)$$

$$\left(-\infty \quad , \quad \bar{x} + t_{1-\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}\right)$$

Justification

- Suppose we have a *normal* population with mean of 3 and standard deviation of 1.
- We take multiple random samples from the population with sample size of 100.
- For each random sample, we construct a $(1 - \alpha) \times 100\%$ confidence interval by following Equation 8.

Justification (Code)



Hypothesis Tests

1. Set up null and alternative hypothesis;
2. Calculate test statistic;
3. Find p value;
4. Make decision;
5. Draw conclusion.

Set Up *Null* and *Alternative* Hypothesis

This is an informative step. We tell people what we are going to on.

1. $$H_0 : \mu < \mu_0 \quad v.s \quad H_A : \mu \geq \mu_0$$

2. $$H_0 : \mu > \mu_0 \quad v.s \quad H_A : \mu \leq \mu_0$$

3. $$H_0 : \mu = \mu_0 \quad v.s \quad H_A : \mu \neq \mu_0$$

We define μ_0. μ_0 can be understood as the value we question about.

Calculate Test Statistic

Test statistic tells us how far the actuality is away from our assumption (μ_0).

$$Z = \frac{\bar{x} - \mu_0}{\sqrt{\frac{\sigma^2}{n}}} \tag{10}$$

In practice,

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{s_X^2}{n}}} \tag{11}$$

Note that $\sqrt{\frac{\sigma^2}{n}}$ (or $\sqrt{\frac{s_X^2}{n}}$) is the standard deviation of sample mean.
Here, our focus is on $\bar{X}$, rather than X.

Find *p*-value

Use tables to find *p*-value.

Make Decisions

- If the test statistic drops in the *rejection region*, p-value $< \alpha$, *reject* H_0;
- if the test statistic does not drop in the *rejection region*, p-value $> \alpha$, *fail to reject* H_0.

We NEVER *accept* H_0.

References I

Waiting times and list statistics: hospitals, Jan 2013.

URL http://webarchive.nationalarchives.gov.uk/20130104155640/http://www.dh.gov.uk/en/Publicationsandstatistics/Statistics/Performancedataandstatistics/HospitalWaitingTimesandListStatistics/index.htm.

R. W. Amin and K. Li.

The Effect of Autocorrelation on the EWMA Maxmin Tolerance Limits.
Journal of Statistical Computation and Simulation, 72(9):719–735, 2002.

J. D. Cryer and K.-S. Chan.

Time Series Analysis With Applications in R.
Springer, 2008.