ECO 602 Analysis of Environmental Data

FALL 2019 – UNIVERSITY OF MASSACHUSETTS

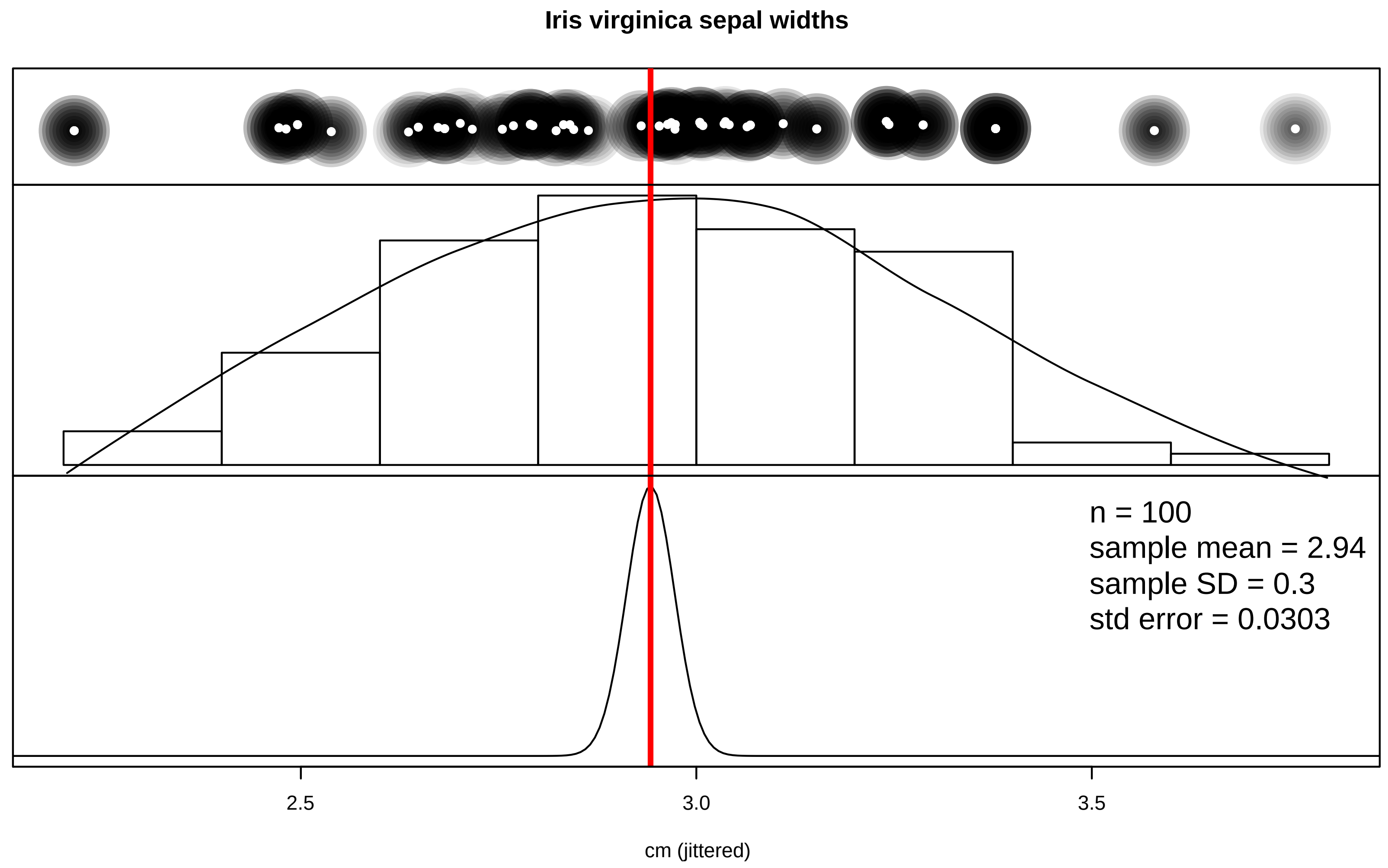
DR. MICHAEL NELSON

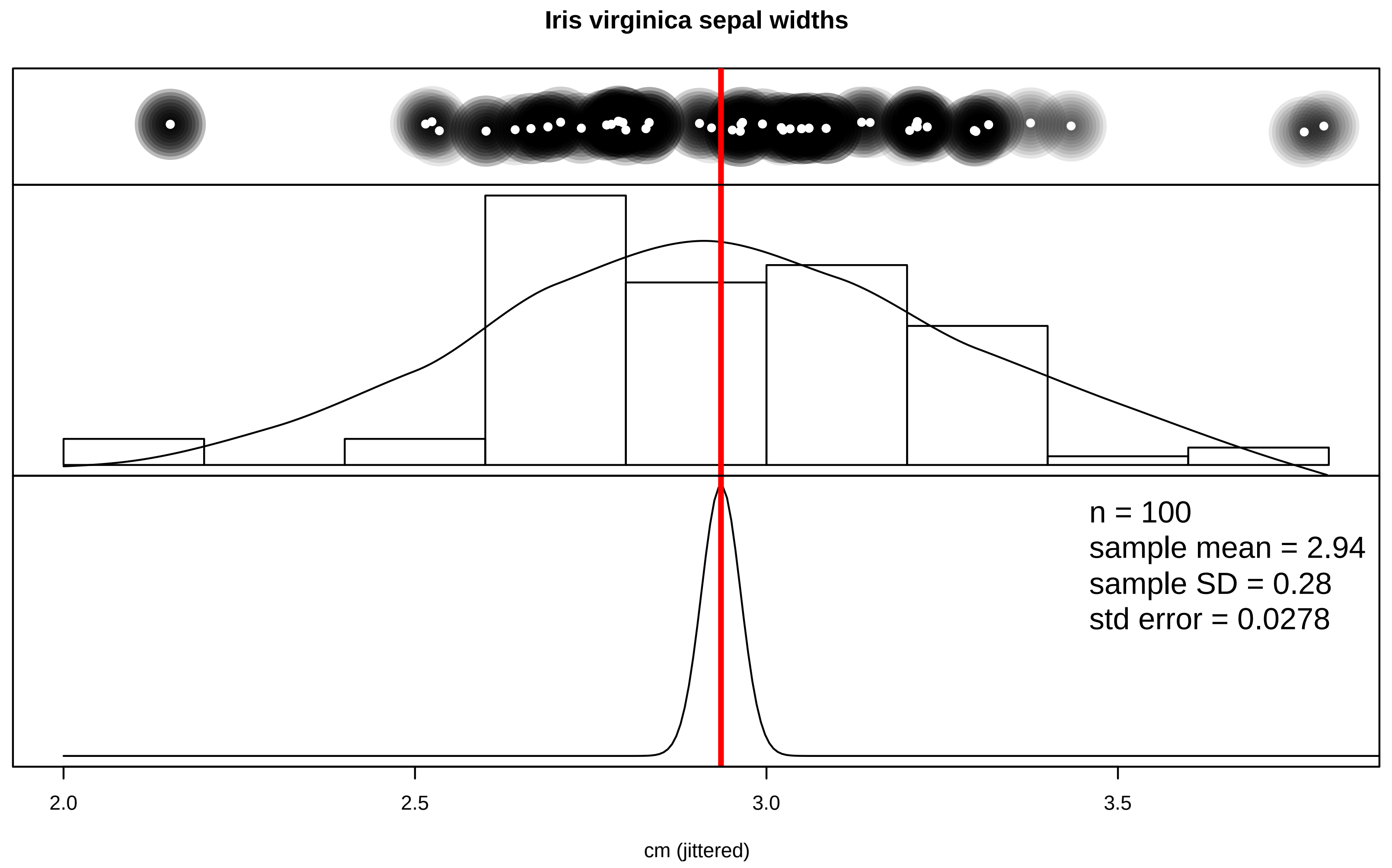
Today's Agenda

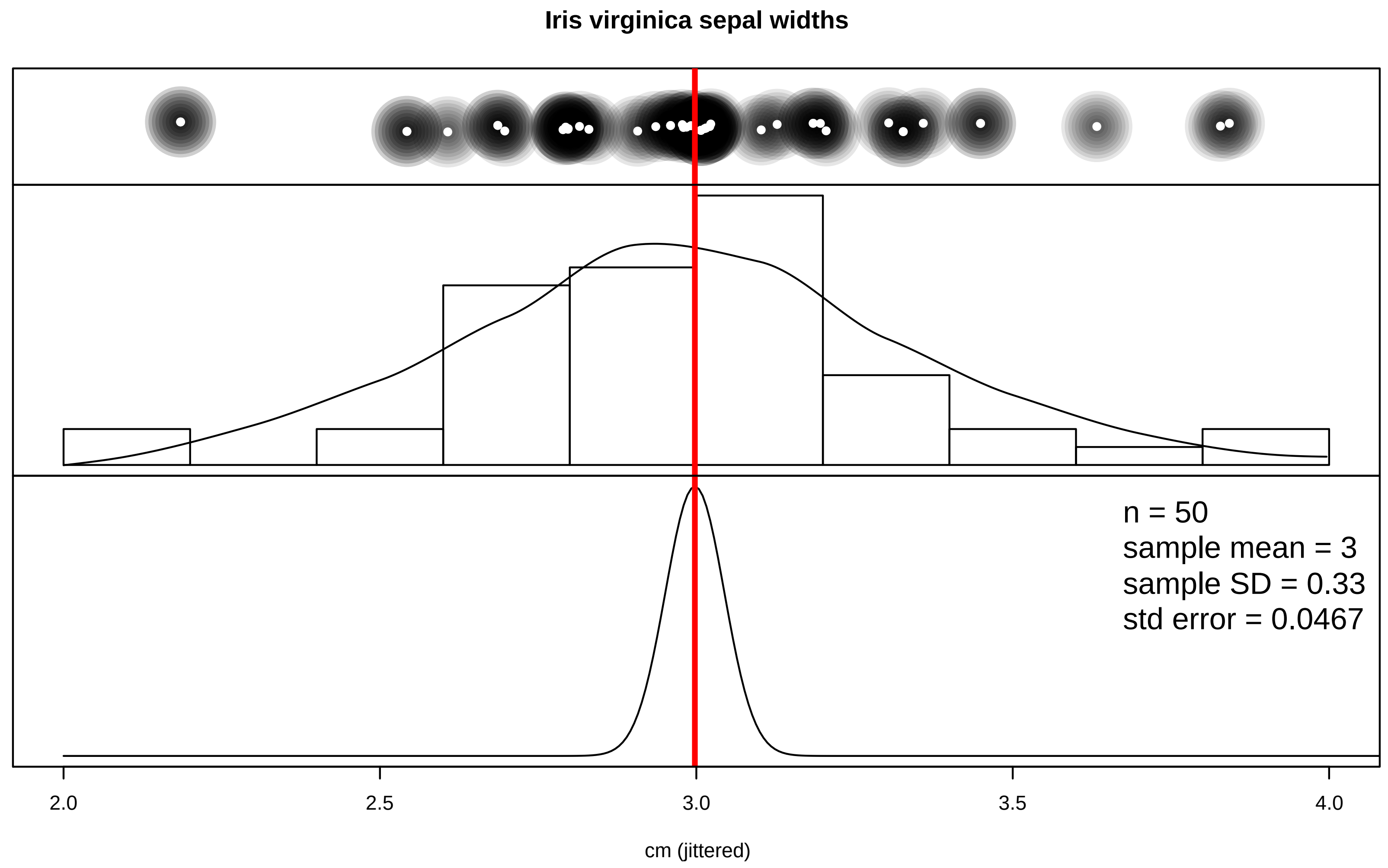
1. Confidence intervals
2. Assignment 3 time
3. Nonparametric OLS inference

Confidence Intervals: Key Terminology

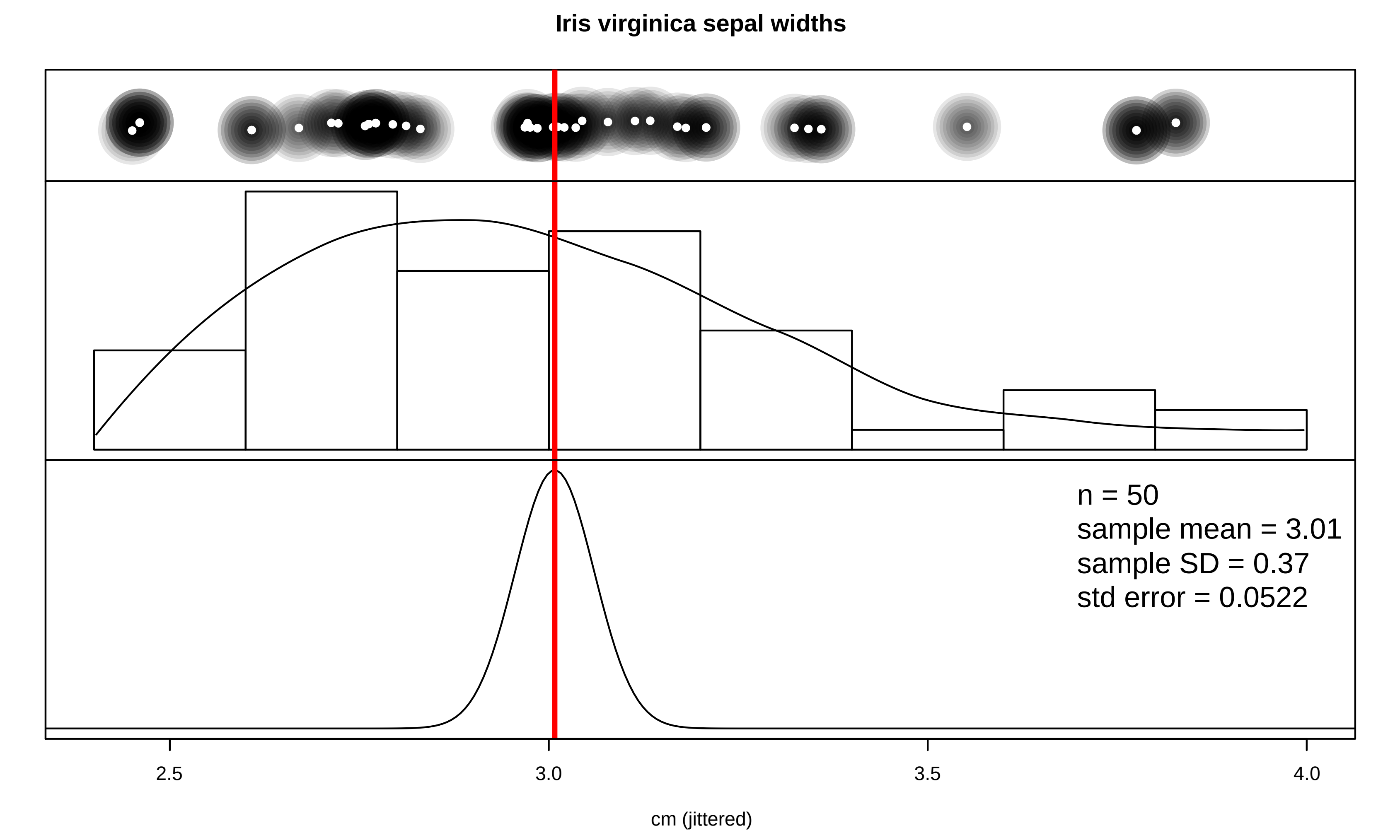
1. Sampling distribution
2. Distribution of sample
3. Sample standard deviation
4. Standard error
5. Standard error of the mean
6. Alpha, beta, statistical power

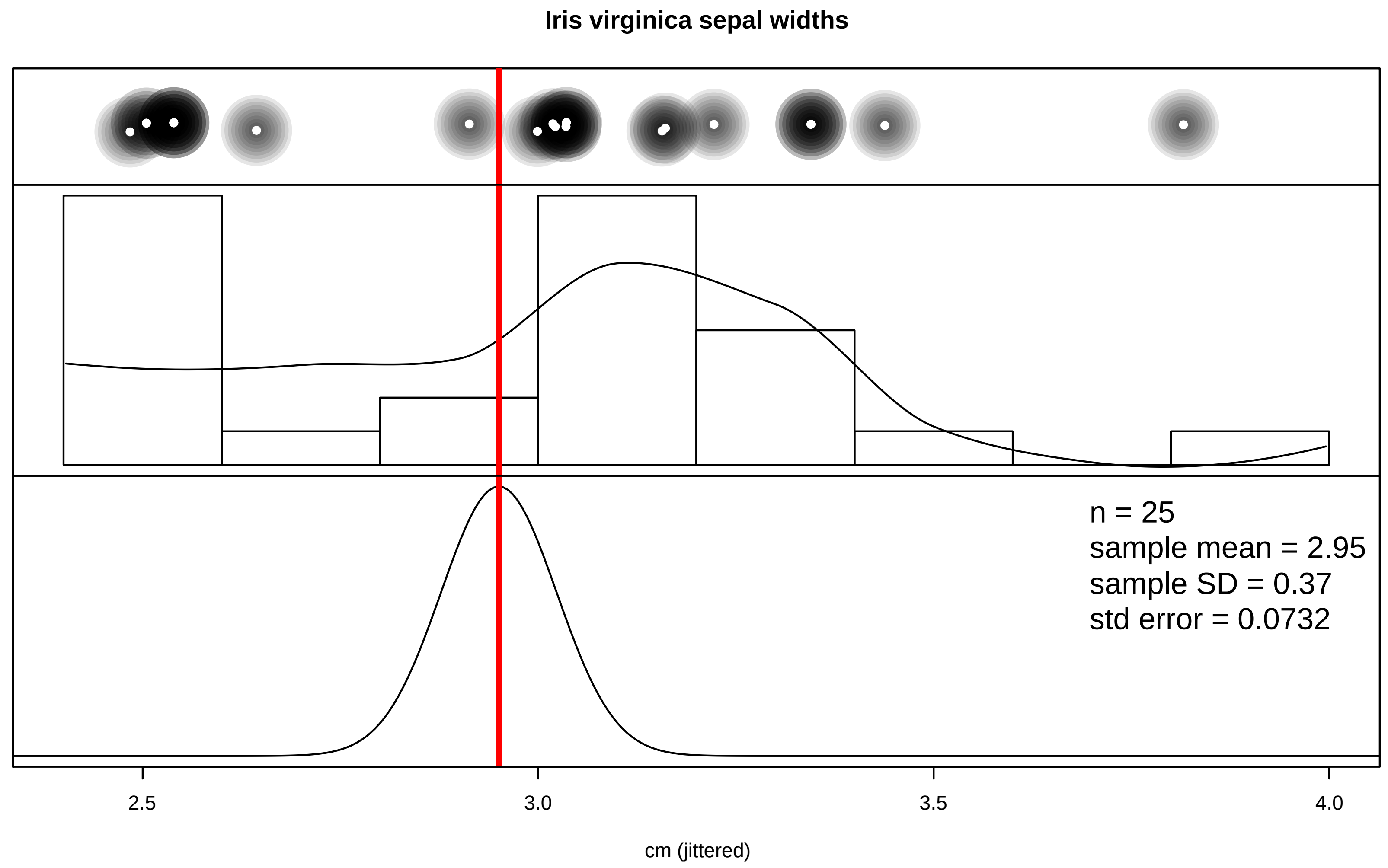

Iris virginica sepal widths
n = 100
sample mean = 2.94
sample SD = 0.3
std error = 0.0303
2.5
3.0
3.5
cm (jittered)


Iris virginica sepal widths
n = 100
sample mean = 2.94
sample SD = 0.28
std error = 0.0278
2.0
2.5
3.0
3.5
cm (jittered)

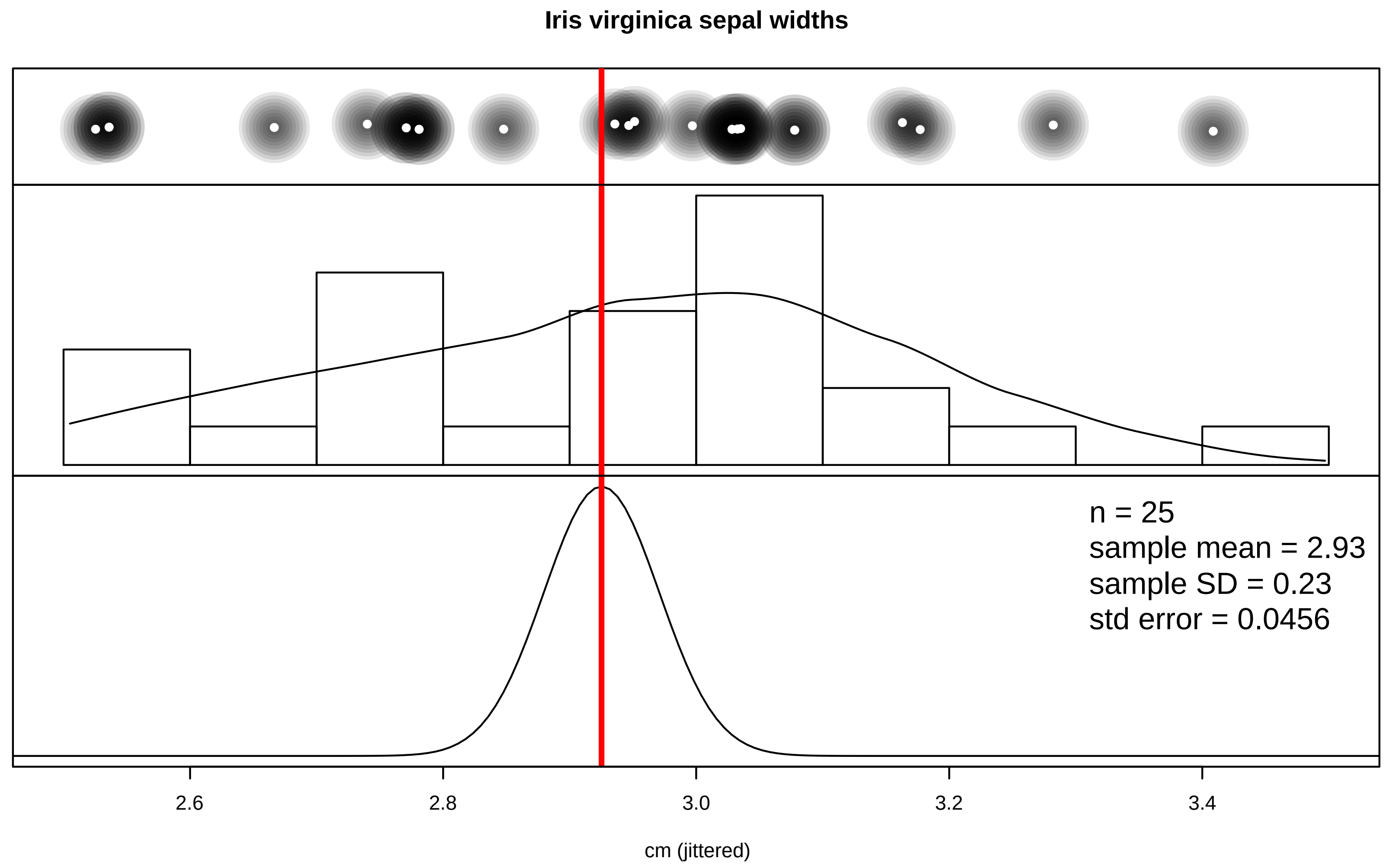


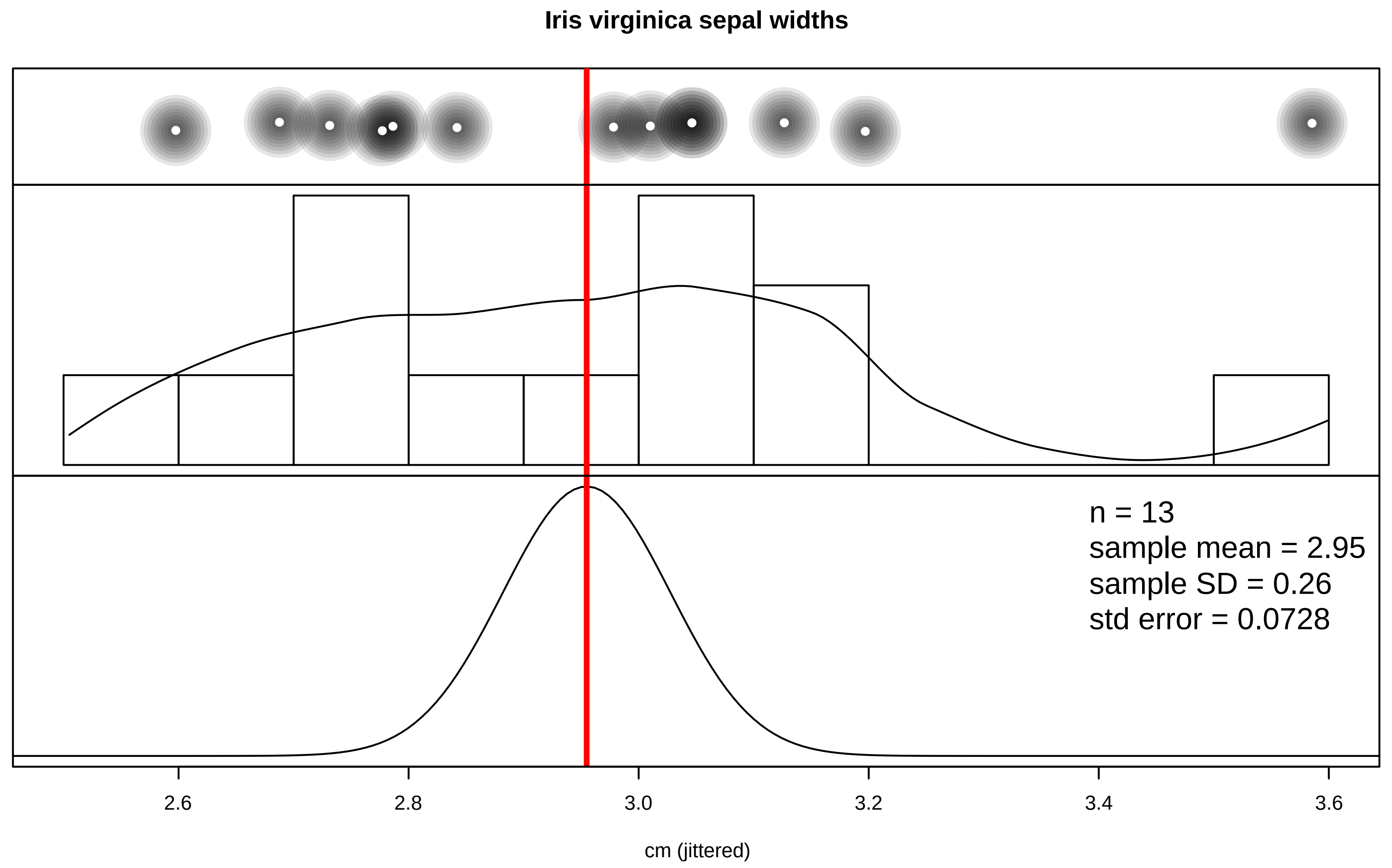
Iris virginica sepal widths
n = 50
sample mean = 3
sample SD = 0.33
std error = 0.0467
2.0
2.5
3.0
3.5
4.0
cm (jittered)

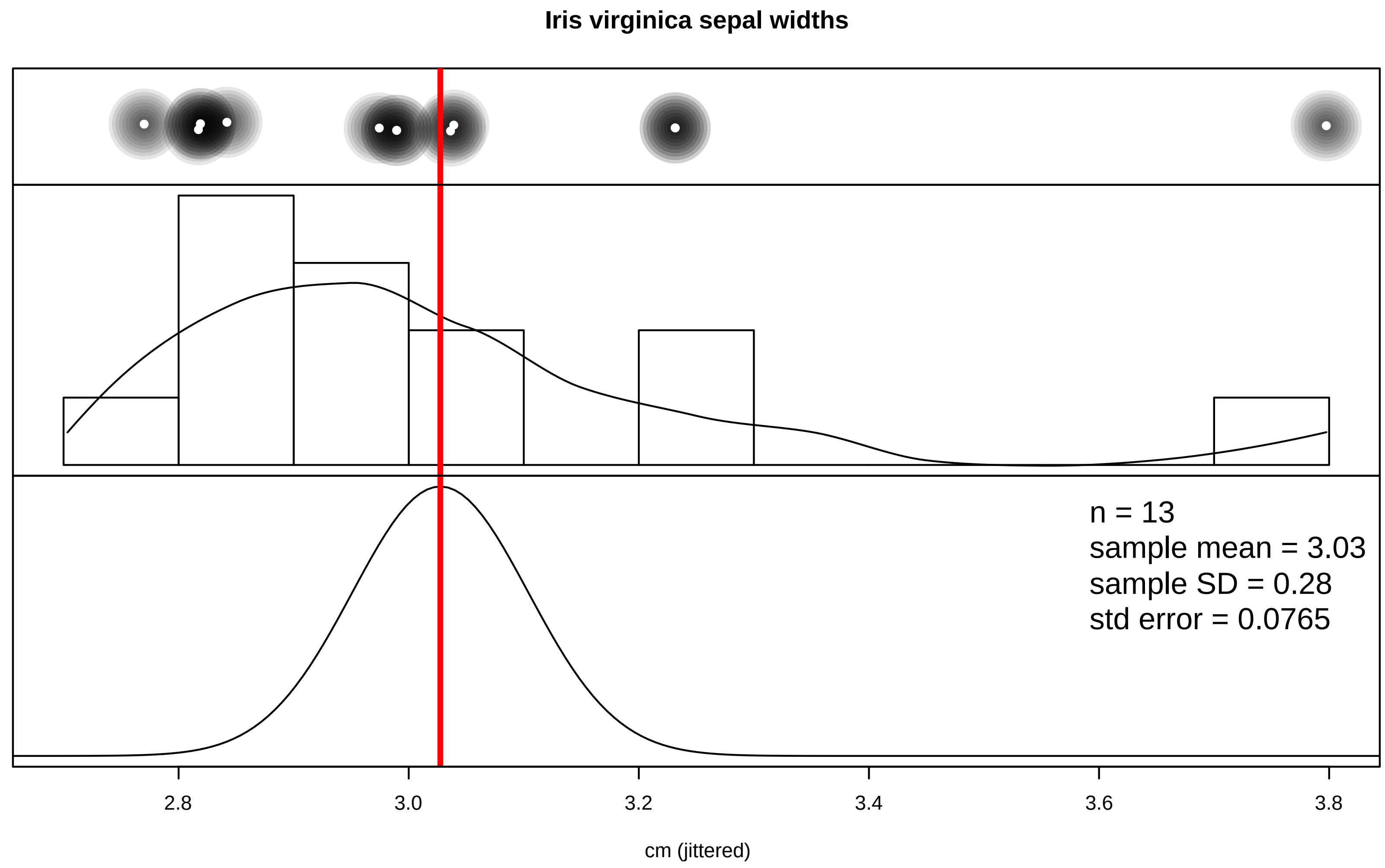

Iris virginica sepal widths
n = 50
sample mean = 3.01
sample SD = 0.37
std error = 0.0522
2.5
3.0
3.5
4.0
cm (jittered)



Iris virginica sepal widths
n = 25
sample mean = 2.95
sample SD = 0.37
std error = 0.0732
2.5
3.0
3.5
4.0
cm (jittered)


Iris virginica sepal widths
n = 25
sample mean = 2.93
sample SD = 0.23
std error = 0.0456
2.6
2.8
3.0
3.2
3.4
cm (jittered)


Iris virginica sepal widths
n = 13
sample mean = 2.95
sample SD = 0.26
std error = 0.0728
2.6
2.8
3.0
3.2
3.4
3.6
cm (jittered)



Iris virginica sepal widths
n = 13
sample mean = 3.03
sample SD = 0.28
std error = 0.0765
2.8
3.0
3.2
3.4
3.6
3.8
cm (jittered)

Things to note:

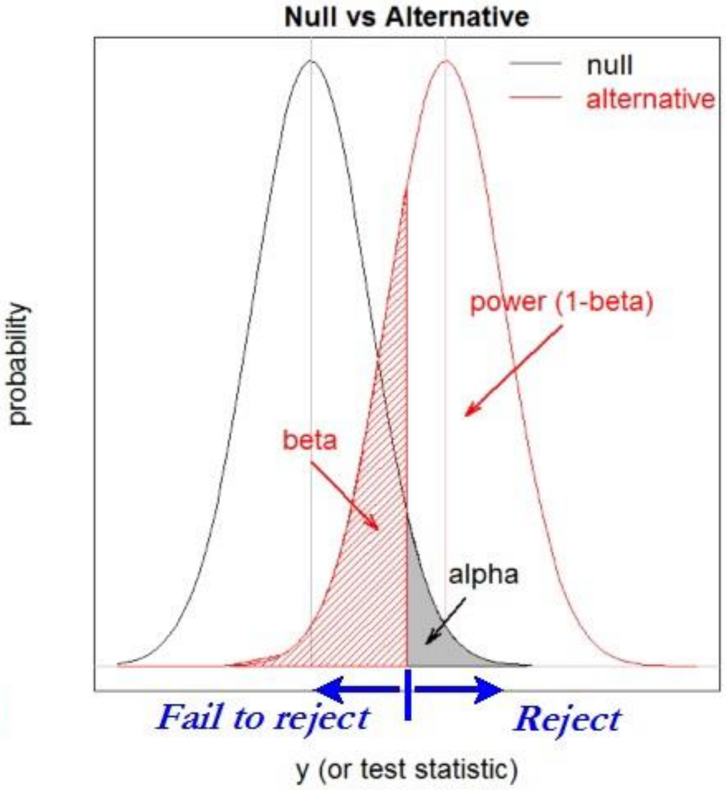
1. Standard error decreases as sample sizes increase.
2. Estimators of population parameters (sample mean, sample variance) stabilize with bigger samples.

Hypothesis Testing Concepts

Neyman-Pearson decision framework

- *alpha* = probability of wrongly rejecting the null hypothesis (Type I error)
- *beta* = probability of wrongly accepting the null hypothesis (Type II error)
- *power* = probability of correctly rejecting the null hypothesis

alpha is under the null; *beta* and *power* are under the alternative



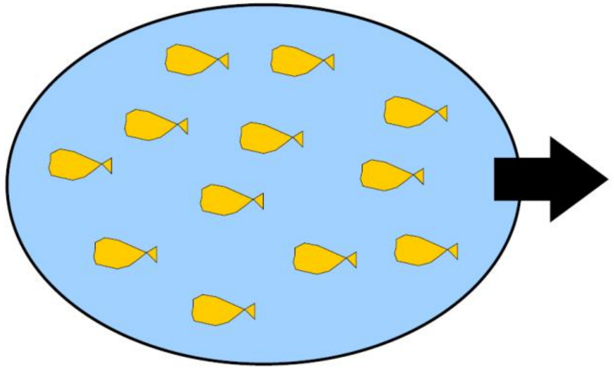
Take a picture of your quiz question answers and email them to me.

Now we're ready to talk about confidence intervals!

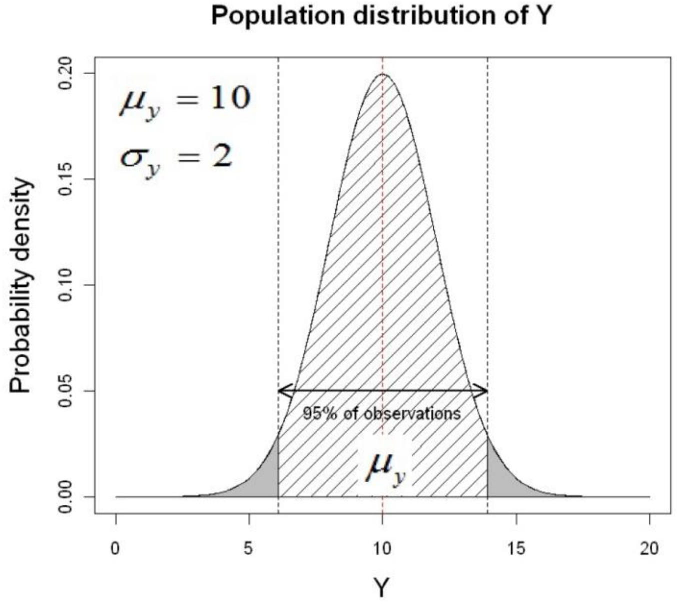
Primer on confidence intervals and more…

Population distribution of a random variable

Population of fish



Y = fish size



$$\Pr\left\{\mu_y - 1.96\sigma_y \le Y \le \mu_y + 1.96\sigma_y\right\} = 0.95$$

This is <u>not</u> a confidence interval!

Primer on confidence intervals and more…

Confidence interval for the sample estimate of population parameter

Confidence interval for the mean:

- Convert the distribution of *sample means* into a standard normal distribution via the *z*-score standardization

σ_y = population standard deviation

$$\sigma_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}} \qquad z = \frac{\bar{y} - \mu_y}{\sigma_{\bar{y}}}$$

$$\Pr\left\{\bar{y} - 1.96\sigma_{\bar{y}} \leq \mu_y \leq \bar{y} + 1.96\sigma_{\bar{y}}\right\} = 0.95$$

This <u>is</u> a confidence interval!

Confidence Interval Demo in R

Was that a letdown?

- Confidence interval: The ‘confidence’ refers to the interval, not the population parameter
- The width of the confidence interval depends on:
  - Alpha
  - Population variability
  - Sample size

Confidence interval is a very frequentist concept.

- Based on hypothetical repeated sampling.
- With alpha = 0.05:
- "If we repeated our sampling scheme many times, around 95% of our confidence intervals would bracket the true population mean."

Confidence interval is a very frequentist concept.

- We can’t say that we are 95% sure a particular CI contains the true population mean.
- A CI either contains the true mean, or it doesn’t… But we cannot tell a particular CI because the true population mean is unknowable.

Nonparametric Inference

- Much of the mathematical hardware is similar to parametric inference.
- Main difference: no attempt to guess a theoretical distribution for the population.
- Main consequence: weaker inference
- ‘Nonparametric’ refers to the lack of an explicit stochastic model for the population.
- We usually calculate statistics in nonparametric inference!

Landscape of Statistical Methods...

The basic statistical model:

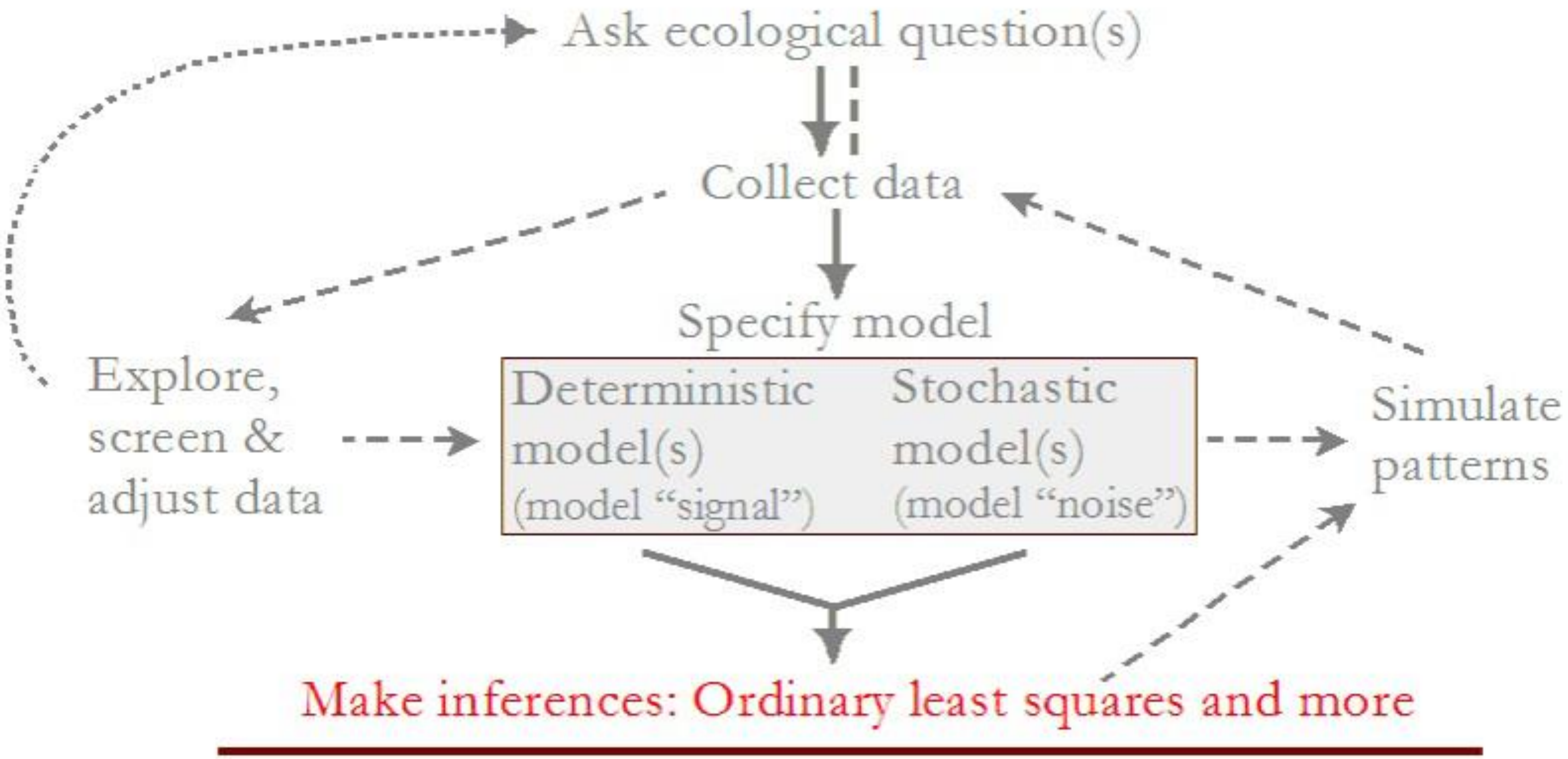
$$\underline{Y} = \underline{\text{deterministic part}} + \underline{\text{stochastic part}}$$

- $\boxed{\text{Univariate}}$
- Multivariate

- $\boxed{\text{Linear}}$
- Nonlinear
- Smoothed

- $\boxed{\text{Distribution}}$
- Heterogeneity
- Autocorrelation
- Multiple levels
- Random noise

Nonparametric Inference



- *Nonparametric inference* involves confronting the model with data to estimate parameters, test hypotheses, compare alternative models, or (with difficulty) make predictions, without specifying a probability distribution

Nonparametric Inference...

Estimate model parameters: OLS method

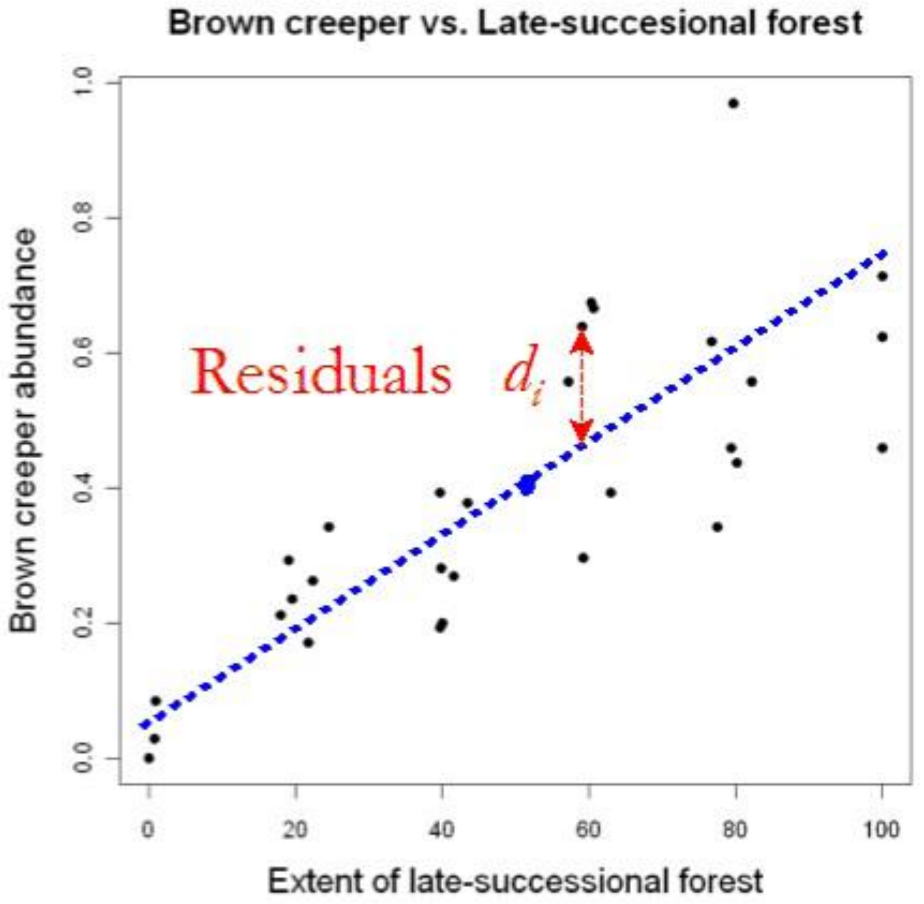
1. Define measure of (lack of) fit:

$$d_i = y_i - \hat{y}_i$$

$$\hat{y}_i = b_0 + b_1 x_i$$

$$d_i = y_i - b_0 - b_1 x_i$$

$$L(Y_i | b_0, b_1) = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$



Likelihood is quantified by minimizing squared errors

Does this sound familiar?

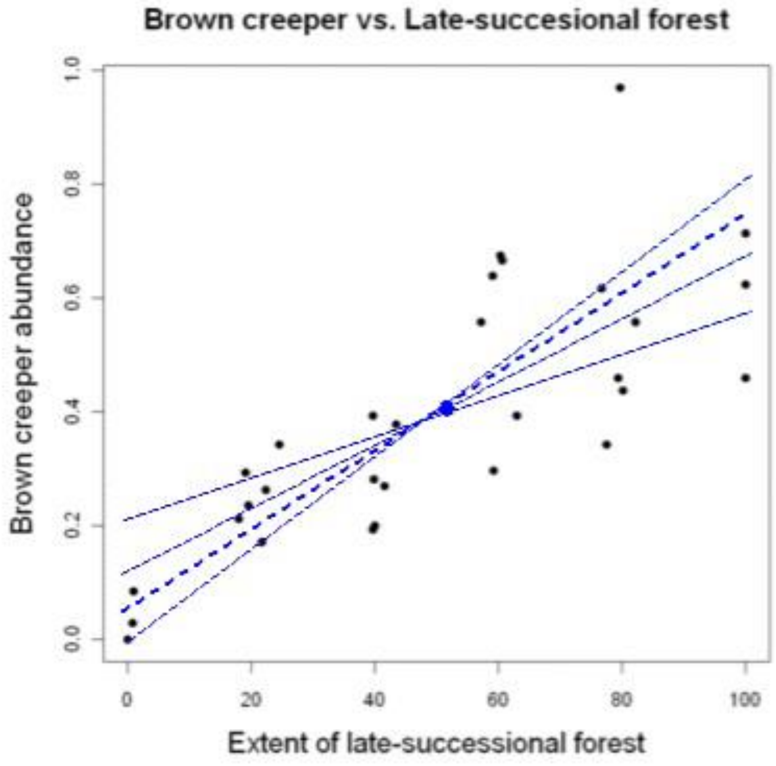
Likelihood function - kind of like Maximum Likelihood Estimation!

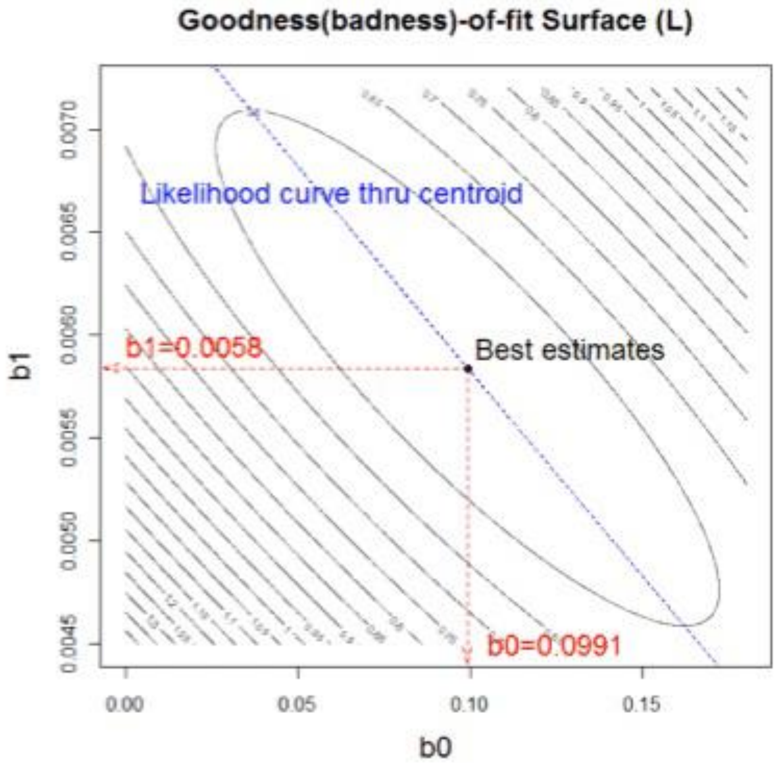
Nonparametric Inference...

Estimate model parameters: OLS method

2. Find estimates that minimize $L(Y_i | b_0, b_1)$

- Numerical solution





Nonparametric Inference...

Estimate model parameters: OLS method

Pros and Cons of OLS Estimation:

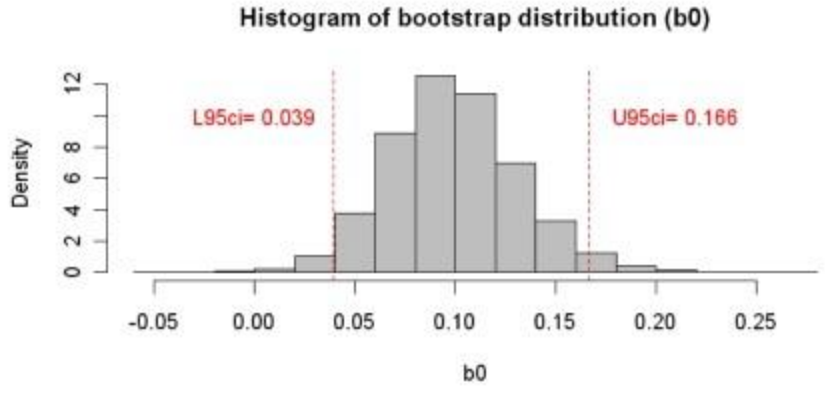
- No assumptions about the error required
- Squared deviations make analytical solutions easier
- If the errors are normally distributed, then the sums of squares is identical to other methods of estimation
- No a priori justification for using the squared measure of deviation, which has an accelerating penalty

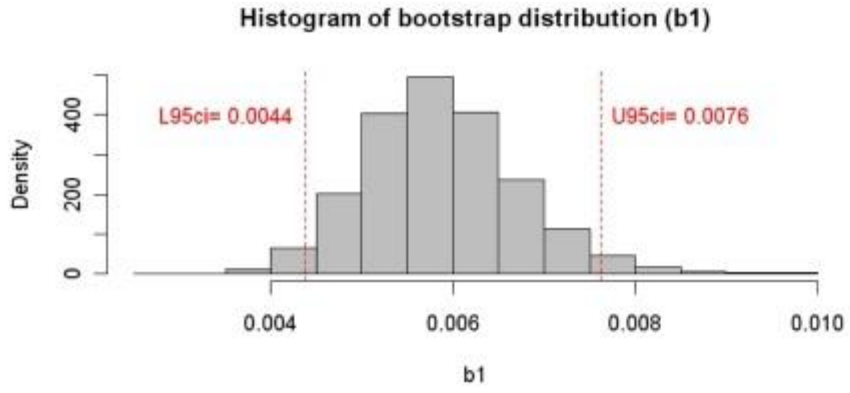
Nonparametric Inference...

Confidence intervals for model parameters

Nonparametric bootstrap confidence interval:

- Repeated sampling of the data, <u>with replacement</u>, to empirically generate the sampling distribution of the estimate
- Quantiles of the bootstrap distribution give the specified confidence interval





What if we are willing to assume something about the population?

Hmmmmmm……

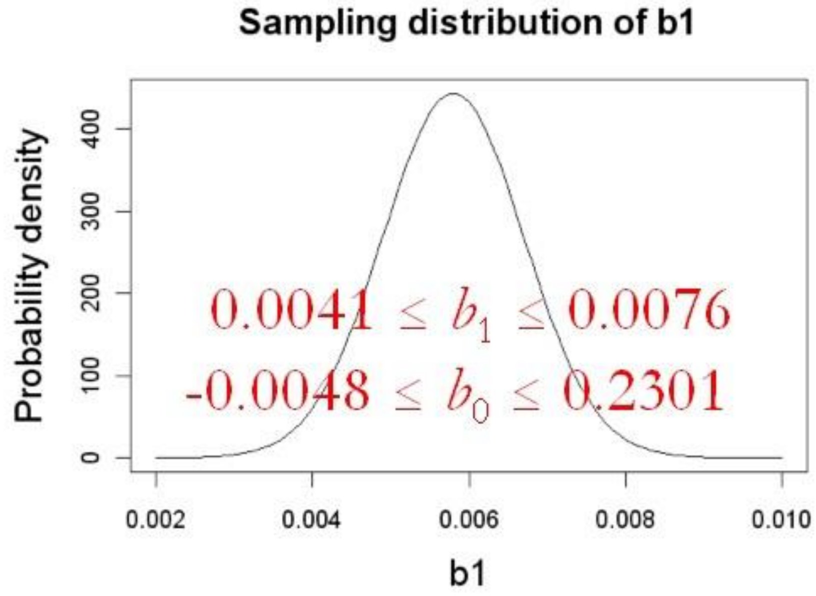
This is sounding more and more parametric…

Nonparametric Inference...

Confidence intervals for model parameters

Parametric confidence interval:

- Calculate the *standard error* of the parameter estimate and multiply it by the appropriate value of Student's *t* and then subtract this interval from, and add it to, the parameter estimate to get the corresponding confidence interval



s^2 = Error variance

$$se_{b1} = \sqrt{\frac{s^2}{SSX}}$$

$$95\%CI = b_1 \pm t_{0.025,n-2}\, se_{b1}$$

$$se_{b0} = \sqrt{\frac{s^2 \sum x^2}{n \cdot SSX}}$$

$$95\%CI = b_0 \pm t_{0.025,n-2}\, se_{b0}$$

For next time:

Keep reading McGarigal chapter 8.

Start reading chapter 9 if you feel adventurous. We'll return to parametric inference after this week.

We will discuss some other nonparametric methods on Thursday.

T-test null and alternative hypotheses

- 1-sample:
- 2-sample:
- Using iris data:
  - 3 species: setosa, virginica, versicolor
  - What are some possible 1-sample hypotheses?
  - What are some possible 2-sample hypotheses?

T-test null and alternative hypotheses

- 1-sample:
- 2-sample:
- Using iris data:
  - 3 species: setosa, virginica, versicolor
  - What are some possible 1-sample hypotheses?
  - What are some possible 2-sample hypotheses?