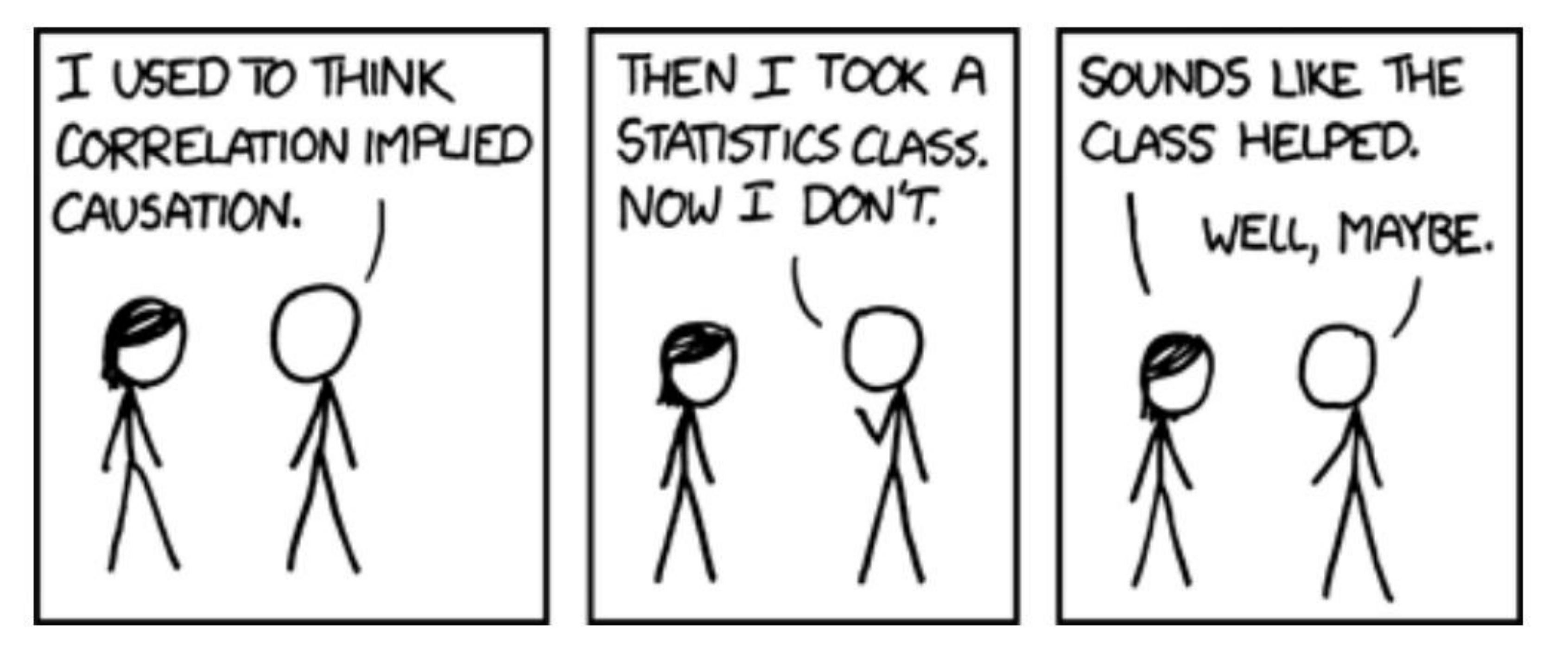
Intro to Statistics Part 2: Estimators, Confidence Intervals, and Bootstrapping

AST1501 Guest Lecture, Nov. 30, 2023

Prof. Gwendolyn Eadie

Slides adapted from what was formerly
“Starfish School” (thank you's to Mubdi Rahman, Renee Hlozek)

Introduction to Basics in Statistics

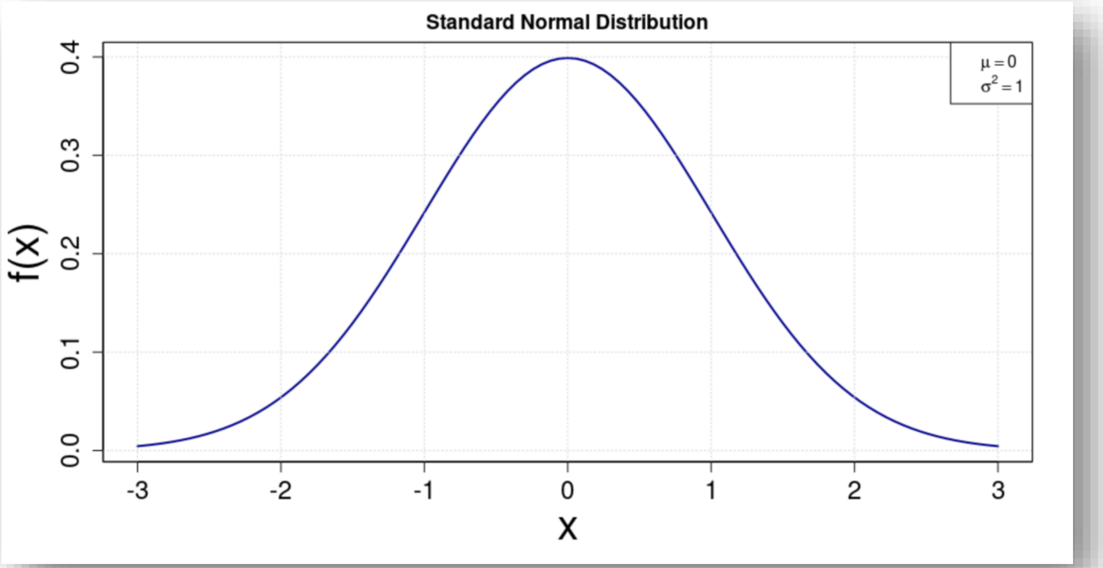


https://xkcd.com/552/

Probability Distributions

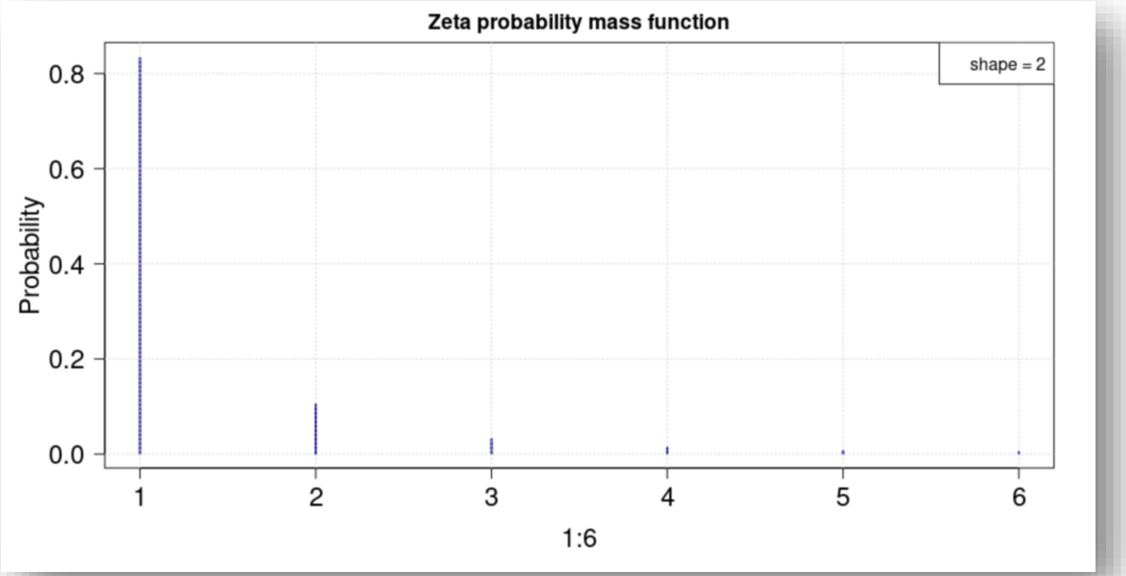
Continuous quantities

probability density function (pdf)

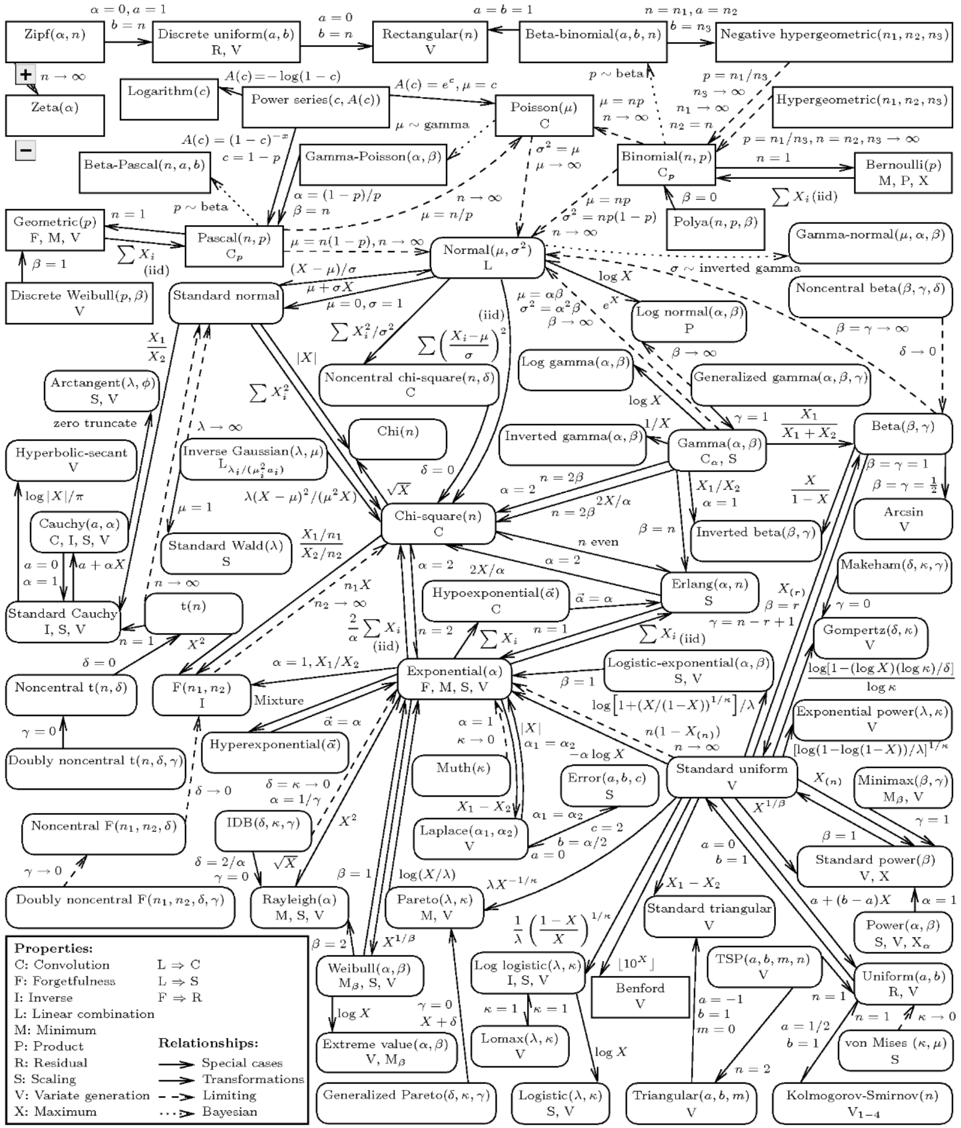


Discrete quantities

Probability mass function (pmf)



There are many univariate distributions!



http://www.math.wm.edu/~leemis/chart/UDR/UDR.html

Confidence intervals

Confidence intervals

- An astronomer has reported that mean log-mass of stars in a globular cluster is $\mathbf{1.30}\boldsymbol{M_\odot}$ with a 95% confidence interval of (1.21, 1.39).
- *What does this interval mean?*

Cool applet to help us understand:

- http://www.rossmanchance.com/applets/ConfSim.html

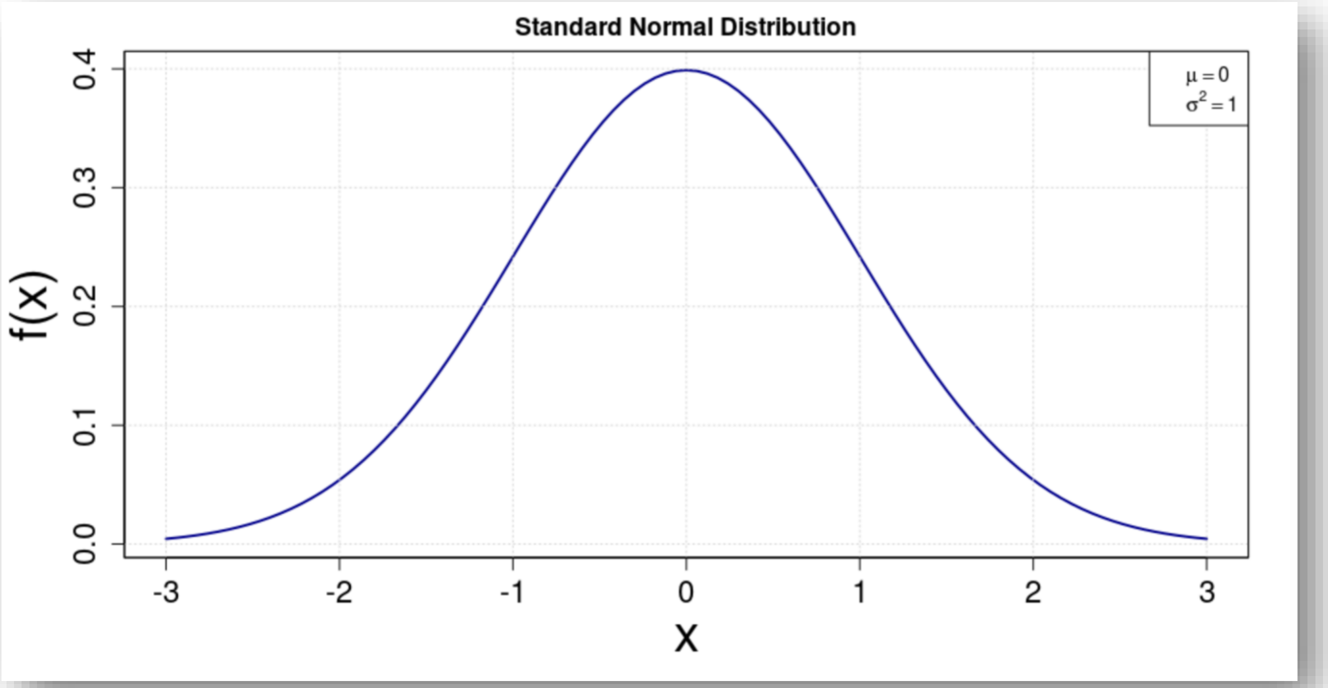
Confidence intervals

- An astronomer has reported that mean log-mass of stars in a globular cluster is $\mathbf{1.30}M_{\odot}$ with a 95% confidence interval of (1.21, 1.39).
- *What does this interval mean?*
  - If you were able to repeat the analysis many times, then 95% of the confidence intervals would overlap the true value. 5% of them would not.
  - In other words, the confidence interval is a random variable!
  - <u>You never know if the confidence interval you calculate contains the true value!</u> It either does, or it doesn't.

- *How do you calculate such a confidence interval?*

Estimators

The Normal Distribution



$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}$$

Estimating parameters

- You think the stars in a globular cluster have log stellar masses that follow a normal distribution.
  - you want to know μ and σ, but you can never know the true values
  - You must **estimate** μ and σ
  - You've collected data on the log masses of 124 stars from the globular cluster
  - You calculate the average of these → estimate of μ
- Estimating the mean

$$\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$\bar{X}$ is an **estimator** of the underlying but unknown population mean μ.

It is a **random variable.**

When you get data, you get a realization of the random variable → $\bar{x}$

Estimators are random variables

- Random variables → follow a distribution
- Example: The estimator $\bar{X} = \sum_{i=1}^{n} \frac{x_i}{n}$ is a random variable that follows a normal distribution:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Note! This is different from the distribution for X, which follows $X \sim N(\mu, \sigma)$

The distribution of the estimator is important → determines the confidence interval

Confidence intervals

Calculating a confidence interval for a mean

- We have the distribution of our estimator for the mean:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Let's standardize it (subtract the mean, divide by the standard deviation):

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

Calculating a confidence interval for a mean

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

Standard Normal:

For a 95% confidence interval, want the lower and upper bounds to be at 2.5% and 97.5% probability

→More generally, for a $\mathbf{1 - \alpha}$ confidence interval:

$$P\left(z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\alpha/2}\right) = 1 - \alpha$$

Calculating a confidence interval for a mean

For a $1 - \alpha$ confidence interval on the mean:

$$P\left(z_{\alpha/2} \le \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le z_{1-\alpha/2}\right) = 1 - \alpha$$

Re-arrange:

$$P\left(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

You can look up the z values in a table (or use software). For a 95% confidence interval on the mean, the z values are ± 1.96

→ The z values will change depending on the confidence interval you want (e.g., 95%, 80%, etc.)

→The lower and upper bound are determined by the distribution of $\bar{X}$

→This assumes *known* population variance σ

Calculating a confidence interval for a mean

When we don't know the population variance σ, we have to estimate it. It's best to use the unbiased estimator

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})^2$$

Now the distribution of the mean is different! Instead of $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$, now we have

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t\ distribution$$

For a $1-\alpha$ confidence interval on the mean, when we don't know the population variance σ:

$$P\left(\bar{X} + t_{\frac{\alpha}{2},n-1}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{1-\frac{\alpha}{2},n-1}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

You can look up the t values in a table (or use software). For a 95% confidence interval on the mean calculated with a sample size $n = 100$, the t values are ± 1.98

→ The t values will change depending on the confidence interval you want (e.g., 95%, 80%, etc.), AND the degrees of freedom $n-1$

→The lower and upper bound are determined by the distribution of $\bar{X}$ (in this case a t-distribution)

Confidence intervals

- An astronomer has reported that the proportion of stars in binary systems is 0.771 with a 95% confidence interval of (0.63, 0.870).
- *What does this interval mean?*
  - If you were able to repeat the analysis many times, then 95% of the confidence intervals would overlap the true value. 5% of them would not.
  - In other words, the confidence interval is a random variable!
  - **<u>You never know if the confidence interval you calculate contains the true value!</u> It either does, or it doesn't.**
- *How do you calculate the confidence interval?*
  - Know the distribution of your estimator
  - Decide on an alpha-level (i.e., what confidence interval % do you want)
  - Look up critical values in tables or compute with software

“3-sigma” detection/error bar in astronomy and physics

And why you should be careful when saying this

The Normal Distribution
(and why astronomers use "sigma")

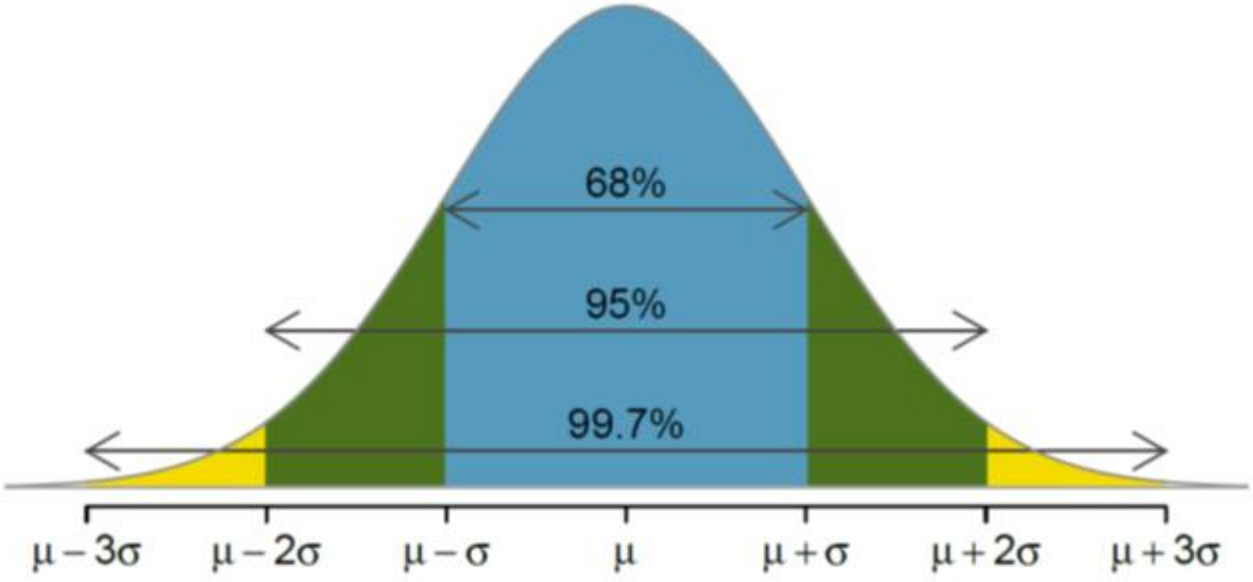
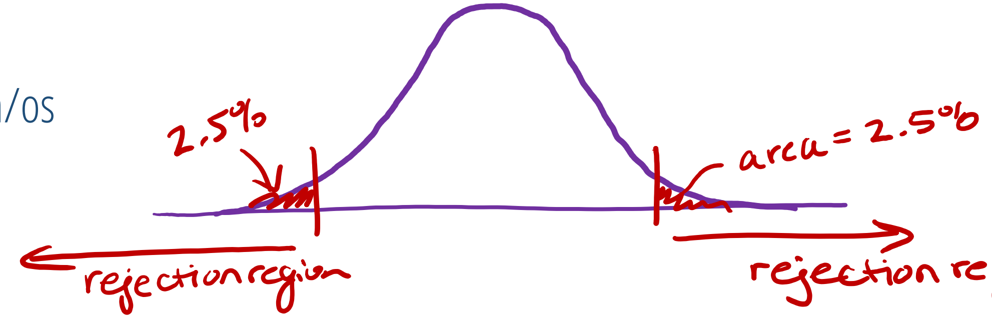


Figure 4.7: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

OpenIntro Stats 4th edition, https://leanpub.com/os



DEFINED BEFORE EXPERIMENT

The Normal Distribution
(and why astronomers use "sigma")

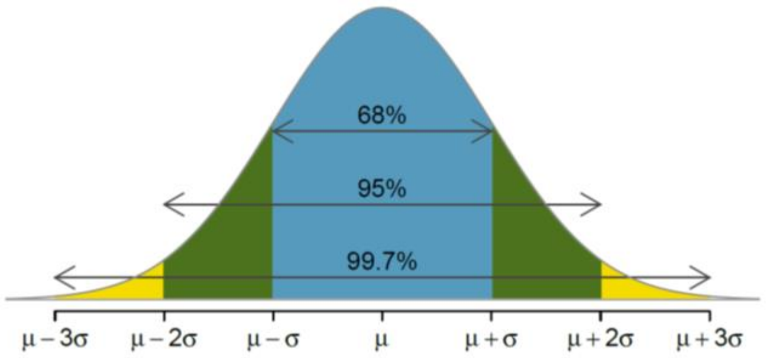
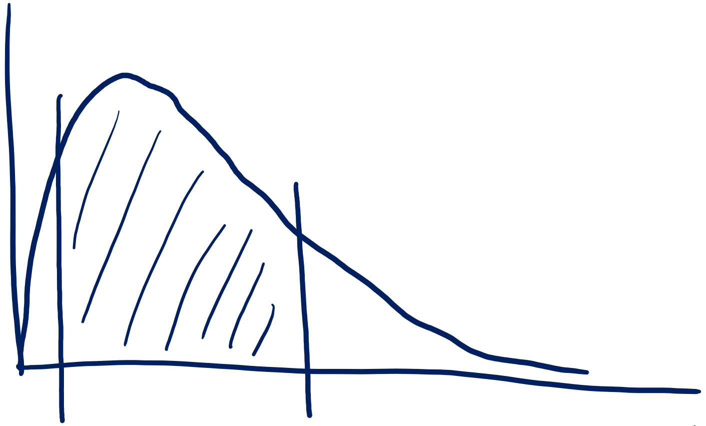


Figure 4.7: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

OpenIntro Stats 4th edition, https://leanpub.com/os



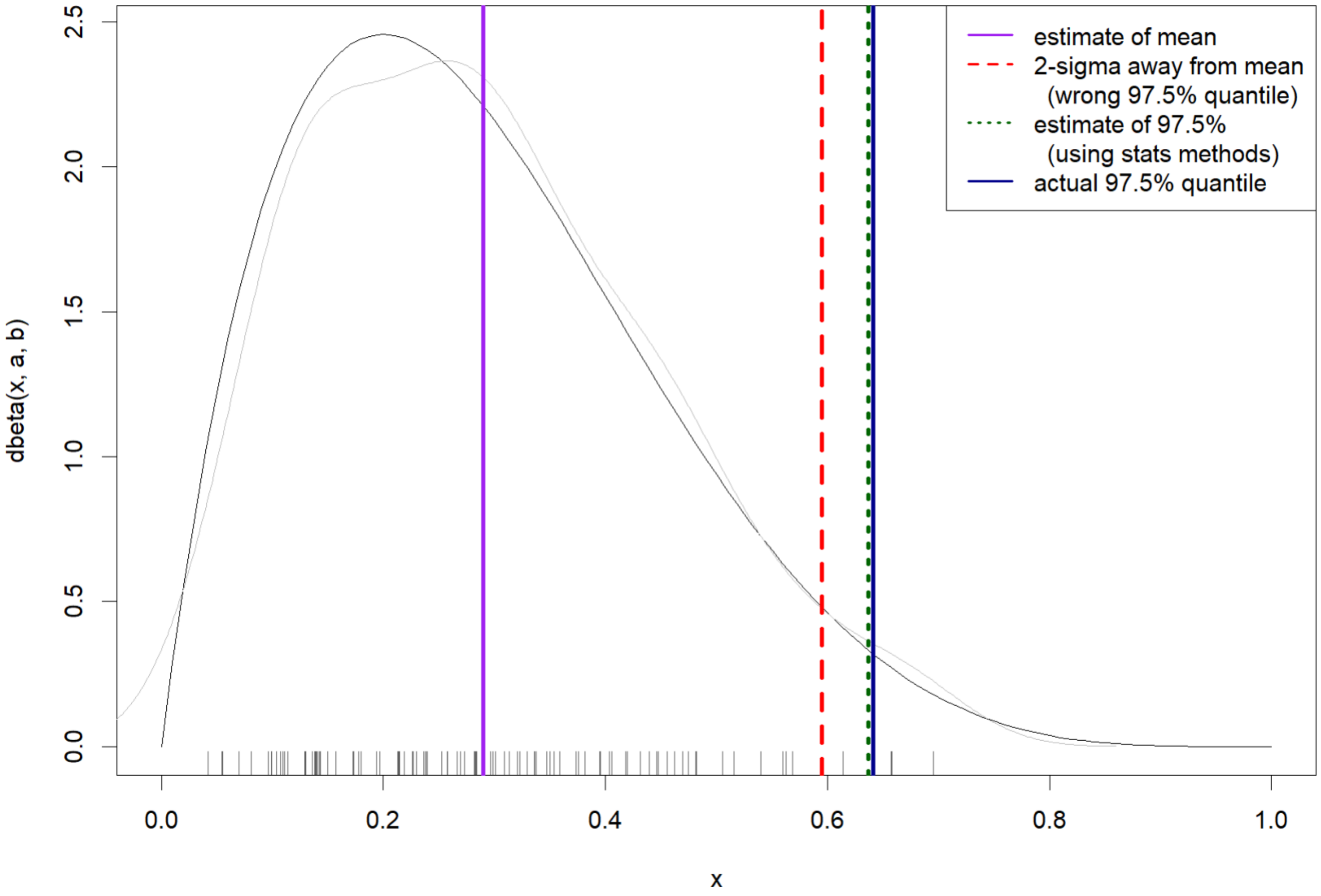
68% → not necessarily = 1sigma

1-sigma = 1 standard deviation

$$\sigma = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

- 2-sigma (2 standard deviations) is equivalent to 95% quantile **only in the case of Normal/Gaussian distributions**
- The 95% quantile is not necessarily equal to 2-sigma in non-Gaussian distributions!

Example where 2-sigma != 97.5% quantile



Underlying distribution:

$$X \sim \text{Beta}(\alpha = 2, \beta = 5)$$

Sampled 100 data points

What if I don’t know the distribution of the estimator?

Bootstrap!

Basics of Bootstrapping

- Create new, mock data sets *using the data set you have*
  - Sample with replacement

- Create many new data sets this way
  - For every data set, estimate the value you are interested in
  - Then look at the distribution of those estimates
  - Calculate the quantile you care about (e.g., inner 90% quantile)

    → this your estimate of the confidence interval

- This works because you are sampling from the empirical distribution

Bootstrapping → when you don't know the underlying distribution

	new (mock) sample #1	new sample #2	many times)
x_1	x_2	x_1	. . . . . .
x_2	x_9	x_8	
x_3	x_3	x_3	
x_4	x_2	x_4	
$\vdots$	$\vdots$	$\vdots$	
x_n	x_3	x_5	
$\downarrow$	$\downarrow$	$\downarrow$	
$\hat{\theta}$	$\hat{\theta}$	$\hat{\theta}$	