Confidence Intervals

DASC 512

Overview

- What is a confidence interval
- How to construct a confidence interval
- What assumptions are required?
- Python application
- Prediction intervals

Standard Error

Recall from earlier, the standard error about the mean:

$$s_{\bar{X}} = \frac{s}{\sqrt{n}} = \sqrt{\left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n(n-1)}\right)}$$

Note that the following hold from this formula:

- Lower population variance → Lower sampling distribution variance
- Larger sample size → Lower sampling distribution variance

Confidence interval

A 100(1-α)% Confidence Interval is a range of values within which the true parameter value will fall for 100(1-α)% of gathered samples.

In other words, if you built equivalent confidence intervals for 100 random samples from the same population, you would expect the true parameter to fall within 95 of those intervals.

This is what it means to be, for example, 95% confident.

Constructing a Confidence Interval

- Define your desired level of confidence $(1 - \alpha)$
- Define the type of confidence interval
  - Symmetric, asymmetric, one-sided
- Define the sampling distribution
  - Any assumptions must be supported by data and sample size

We'll start by looking at symmetric CIs using a normal sampling distribution

Margin of Error

The margin of error or confidence interval half-width for a symmetric two-sided confidence interval is the distance from the point estimate to the edges of the interval

Confidence Interval about the Mean

If we can assume a normal sampling distribution,

$$ci = \bar{x} \pm \text{SEM} \times N_{PPF}\left(1 - \frac{\alpha}{2}\right)$$

where (1-α) is the desired confidence level, SEM is the standard error about the mean, and N_{PPF} is the percentile point function for the standard normal distribution (often called the Z critical value, or z* as in the book)

Note: If using Python, there is rarely a reason to use z*. Use t* instead with $\nu = n - 1$ degrees of freedom. The use of z* makes table lookups easier.

Confidence Interval about the Mean

$$ci = \bar{x} \pm \text{SEM} \times N_{PPF}\left(1 - \frac{\alpha}{2}\right)$$

Given the estimated sampling distribution $\bar{X} \sim N\left(\bar{x}, \frac{s}{\sqrt{n}}\right)$, a (1-α) confidence interval (a,b] is any interval for which

$$P(a < \bar{x} \leq b) = 1 - \alpha$$

Confidence Interval

This concept can be extended, then, to any distribution and parameter of interest. Given parameter of interest $\tilde{X}$ distributed

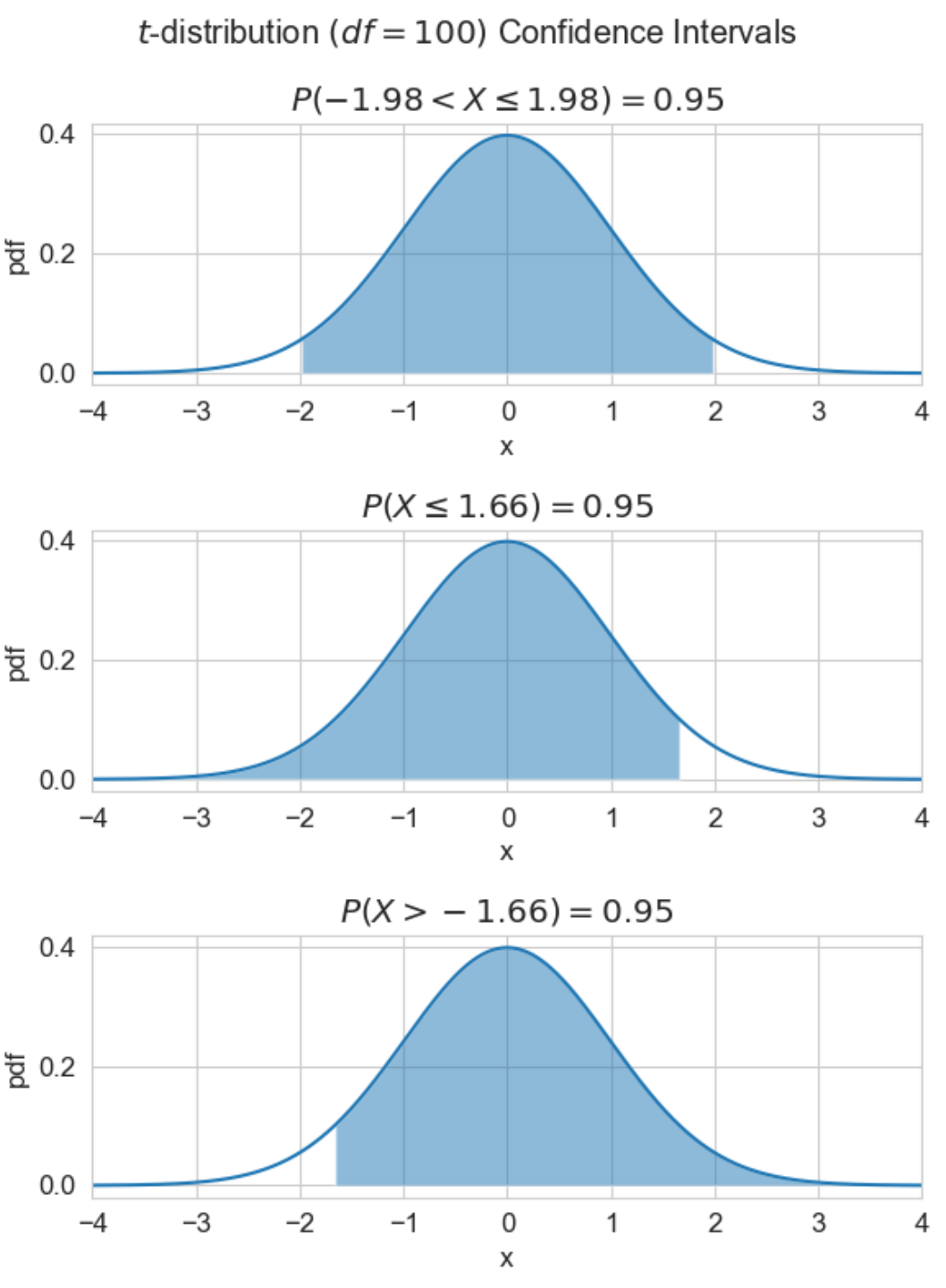
$$\tilde{X} \sim \gamma(\theta)$$

where γ is some distribution and θ are its parameters,

$$ci = \left(\gamma(\theta)_{PPF}\left(\frac{\alpha}{2}\right), \gamma(\theta)_{PPF}\left(1 - \frac{\alpha}{2}\right)\right], \text{or}$$

$$ci = (-\infty, \gamma(\theta)_{PPF}(1-\alpha)], \text{or}$$

$$ci = (\gamma(\theta)_{PPF}(\alpha), \infty]$$



Inverse Functions for Discrete Distributions

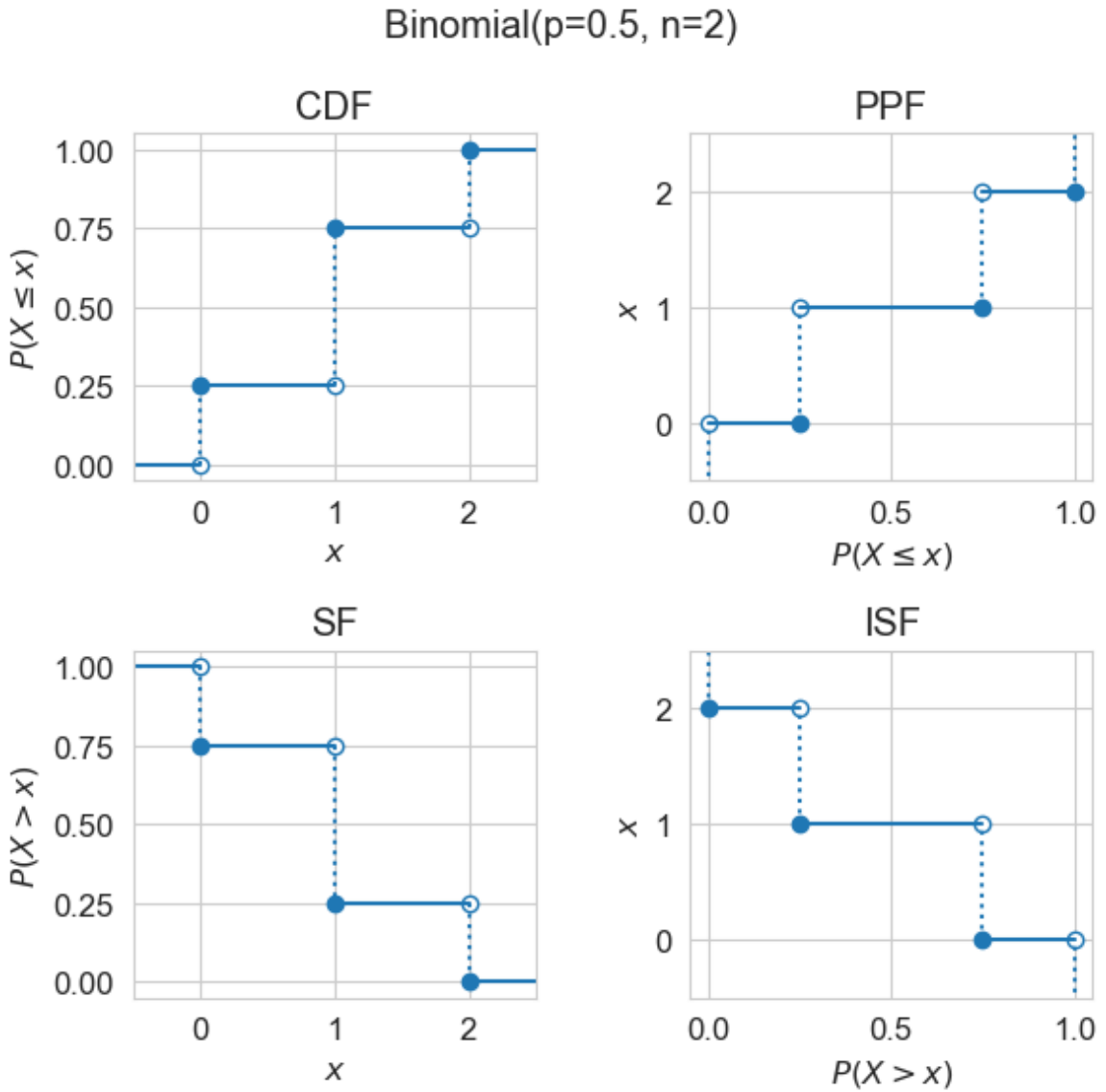
Recall that the interpretations of PPF and ISF vary slightly for discrete RVs

See at right, PPF maps onto dotted lines (e.g., F(x) is never 0.1)

For input quantile q, the output is the most conservative limit allowing that

PPF: $P(X \leq x) \geq q$

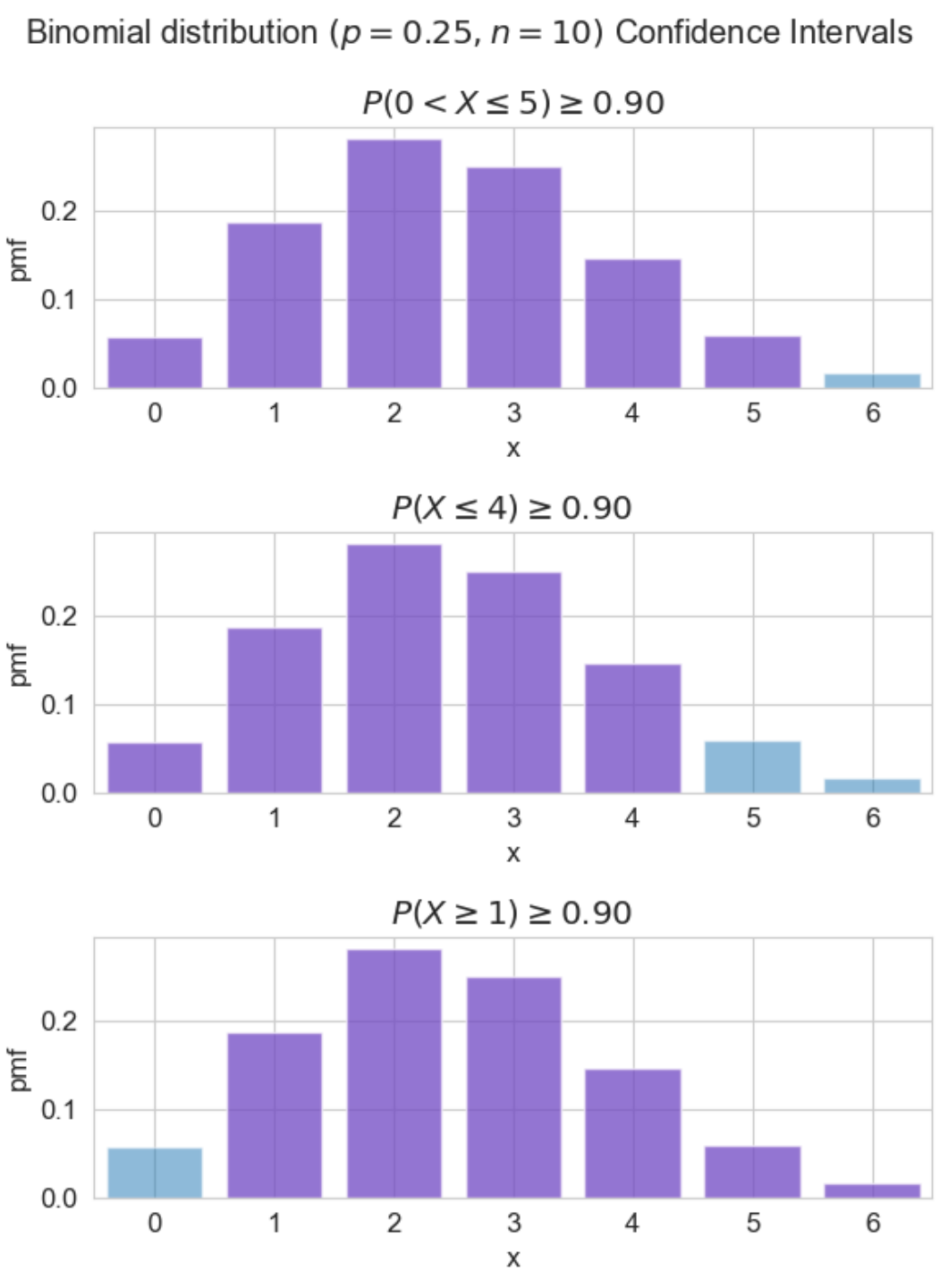
ISF: $P(X > x) = 1 - P(X \leq x) \leq q$



Binomial CIs

The change in interpretation makes our intervals closed rather than right-closed

$$ci = \left[\gamma(\theta)_{PPF}\left(\frac{\alpha}{2}\right), \gamma(\theta)_{PPF}\left(1-\frac{\alpha}{2}\right)\right], \text{or}$$
$$ci = (-\infty, \gamma(\theta)_{PPF}(1-\alpha)], \text{or}$$
$$ci = [\gamma(\theta)_{PPF}(\alpha), \infty)$$



Assumptions

- **Always**: Independent and identically distributed (iid) data collection
- Type of data: numerical or categorical
- Distribution: typically normal or *t* (CLT), but may be another distribution
- Sample size: must be sufficient to justify distributional assumption

Building in Python

- Normal confidence intervals
- *t* confidence intervals
- Binomial confidence intervals

Prediction Intervals

A 100(1-α)% Prediction Interval is a range of values within which the next observation is expected to fall 100(1-α)% of the time.

In other words, if you collected a random sample of 100 new data points from the same population, you would expect 95 of those observations to be within that interval.

Prediction Interval about the Mean

$$ci = \bar{x} \pm \sqrt{\frac{1}{n}} s \times t_{PPF,\nu=n-1}\left(1 - \frac{\alpha}{2}\right)$$

$$pi = \bar{x} \pm \sqrt{1 + \frac{1}{n}} s \times t_{PPF,\nu=n-1}\left(1 - \frac{\alpha}{2}\right)$$

Recap

- What is a confidence interval
- How to construct a confidence interval
- What assumptions are required?
- Python application
- Prediction intervals