Mathematical Foundations for Data Science
DA5000

Session 4 – Inferential Statistics

Nandan Sudarsanam,
Department of Data Science and AI,
Wadhwani School of Data Science and AI,
Indian Institute of Technology Madras

The concept of inferential statistics

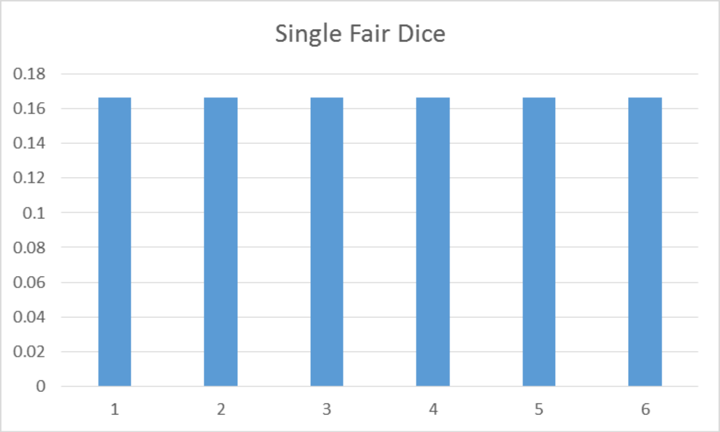
- Descriptive versus Inferential. The use of Sample and Population
  - Population as a bigger data set
  - Population as a phenomena (random variable vs variates)
- Some examples:
  - Marketing: Our discrete example of sales calls
  - Physical systems: Our continuous example of ball drop
  - Operations: The weight of bags of chips
  - Finance: Stock returns/price
- The two-sample (and multiple sample) setting

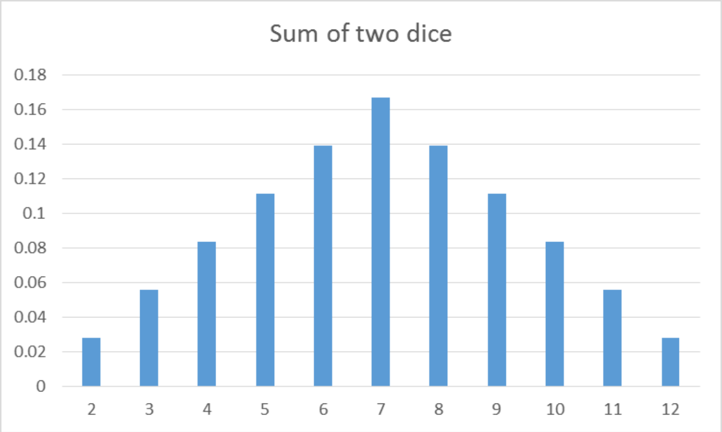
So where can I use it?

- Making an inference about a population from a sample
  - Given a sample statistic ($\bar{x}$) what can I say about the population parameter (μ): Confidence Intervals
  - Given a sample can I answer pointed questions about the population: Hypothesis Tests

Central Limit Theorem

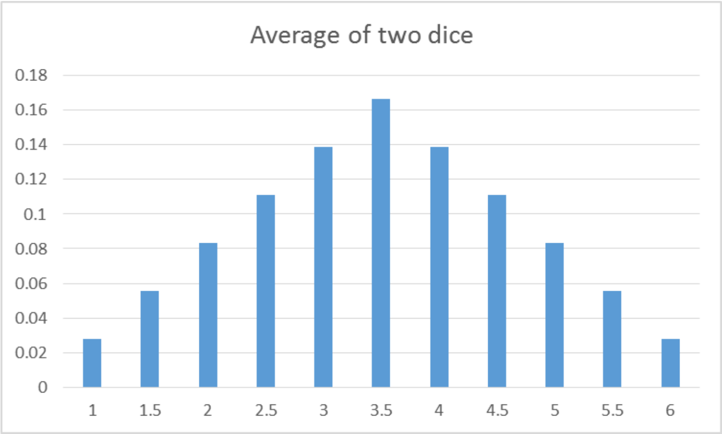
- The aggregation of a sufficiently large number of independent random variables results in a random variable which will be approximately normal.

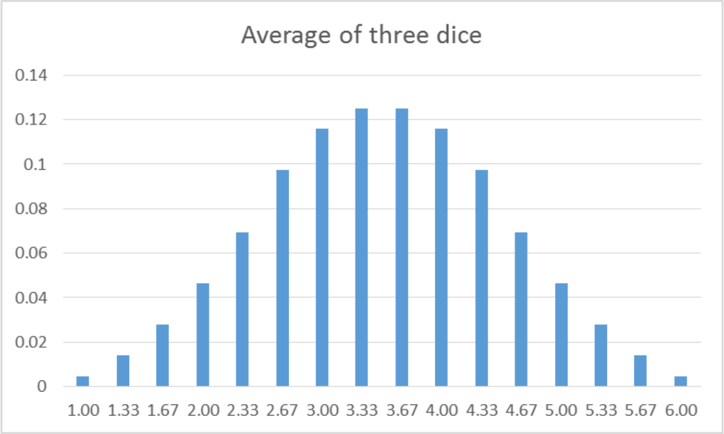




Central Limit Theorem

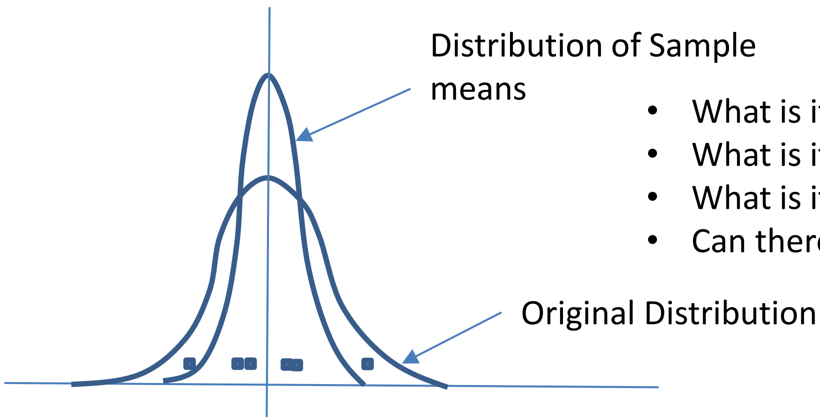
- More distributions:





Sampling distribution

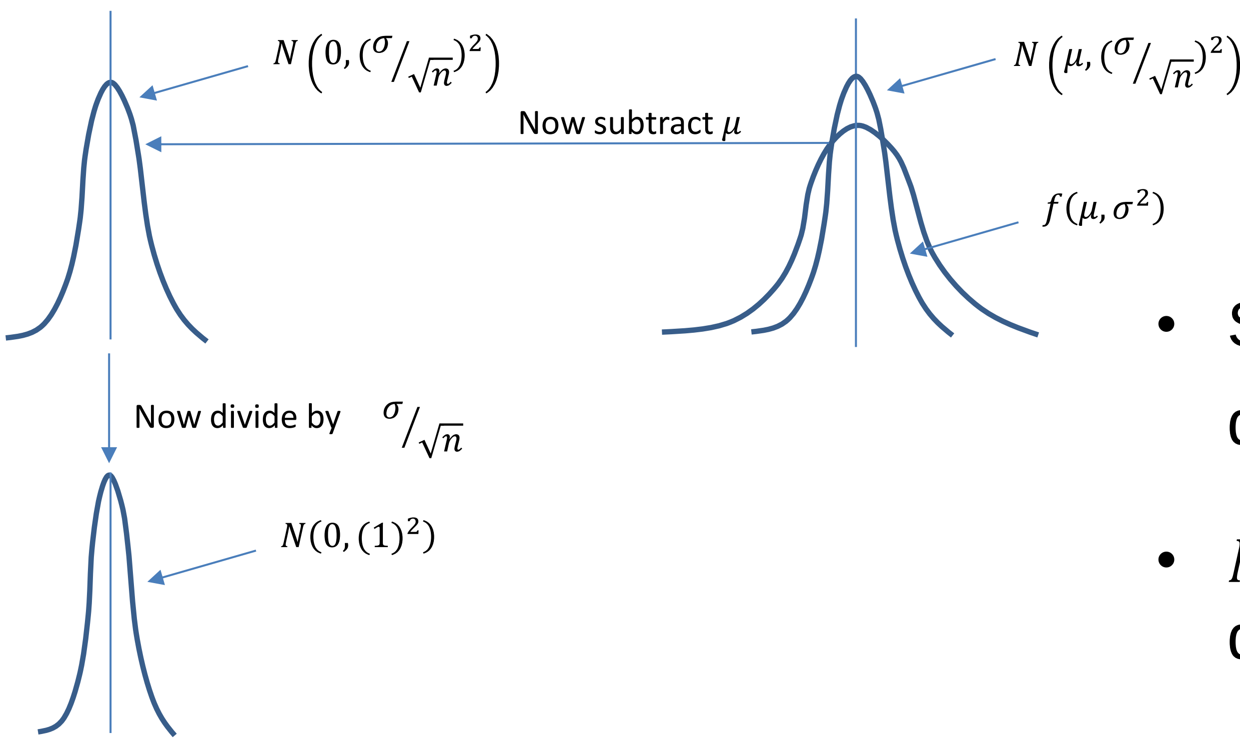
- Sampling distribution



- What is its shape?
- What is its mean?
- What is its standard deviation?
- Can there be a distribution for sample standard deviations?

Single sample interval

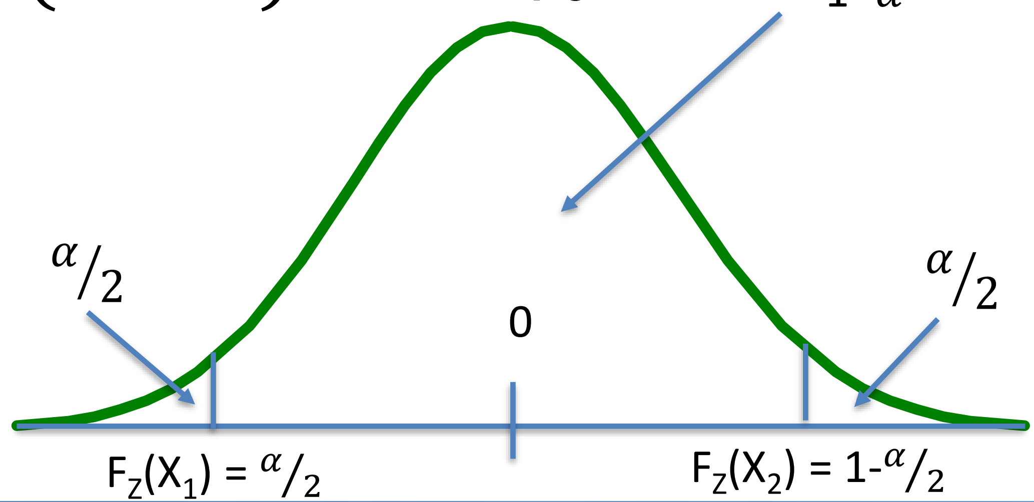
- Idea: if I can look at how much $\bar{X}$ can deviate from μ, then for a given $\bar{x}$, I can quantify the range within which μ could exist.



- So what is the distribution of $\frac{\bar{X}-\mu}{{}^{\sigma}/_{\sqrt{n}}}$
- $N(0, 1^2)$ or Z distribution

Single sample interval for mean

- How far does the $N(0,1^2)$ or Z extend on either side?
- Certainty as percentage: a small chance of an error α, implies that we are certain with a $(1-\alpha) * 100\%$



X_1 is referred to as $-z_{\alpha/2}$ and X_2 is referred to as $z_{\alpha/2}$

Steps

Therefore

$$P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] = 1 - \alpha$$

$$P[-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}] = 1 - \alpha$$

$$-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

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

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

Examples and discussion

- Examples
  - The ball drop example
  - What is the average waiting time for a patient in doctor's office?
  - What is the average diameter of a part we manufacture?
  - What is the average rating (on 10) for a dish in a restaurant
- Concrete steps: What is average weight of a bag of chips? We know $\sigma = 2$gms. We sample 10 bags and find that they weigh 99,100,102,101,100,101,100,99,100,101
- We first find $\bar{x}$ =100.3, then we can say that the 95% confidence interval around μ is ($\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 100.3 \pm 1.96\frac{2}{\sqrt{10}}$ = $\{99.06 \leq \mu \leq 101.54\}$
- One sided bound?

Other confidence intervals

- When variance is unknown: $\bar{x} \pm t_{\alpha/2,n-1} \dfrac{s}{\sqrt{n}}$
- Examples: Same as when variance was known
- Large-sample confidence interval for proportions:
- Examples: Sales calls

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Confidence interval on the variance:

$$\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}$$

- Examples: Variance in the bag of chips