UNIVERSITY of WASHINGTON

Normal Distribution
Normal Approximation

MATH/STAT 394: Probability I
Summer 2021 A Term

Introduction to Probability
D. Anderson, T.Seppäläinen, B. Valkó

§ 3.5, 4.1, 4.3

Aaron Osgood-Zimmerman

Department of Statistics

Logistics

- course evaluation: https://uw.iasystem.org/survey/245028 (*please consider filling this out to improve the course and my teaching for future students!*)
- HW4 due tonight at 11:59pm
- HW5 (last HW) available now, due next Tuesday at 11:59am (noon) PST so we can release HW5 solutions for you to review before the final is due
- Final
  - will primarily cover material since the midterm, but you may need leverage knowledge you learned per-midterm
  - will be available after the final lecture of new material (Monday July 19)
  - will be due Wednesday July 21 at 11:59pm
  - unlimited time allowed during that window
- Last day of lecture will be Q+A (basically extra office hours)
- I have posted a review lecture deck that you can look at beforehand if you like

Practice solution

Practice

Find from a table (on the web or from the book) a value z s.t.

$$\mathbb{P}(-z \leq Z \leq z) = 95/100$$

Solution You can find $z = 1.96$

Outline

Gaussian distribution

Lemma
Let $Z \sim \mathcal{N}(0,1)$, *then* $\mathbb{E}[Z] = 0$ *and* $\mathrm{Var}(Z) = 1$

Proof

- First check that $\mathbb{E}[Z]$ is well defined, which means showing that $\mathbb{E}[|Z|] < +\infty$. For that one shows that $\int_{-\infty}^{+\infty} |x| e^{-x^2/2} dx = 2 \int_0^{+\infty} x e^{-x^2/2} dx = 2$ is finite
- Then since the p.d.f. of Z satisfies $\phi(x) = \phi(-x)$, we have that (ϕ is the p.d.f. of Z)

$$\int_{-a}^{a} \phi(x) dx = 0$$

- Therefore $\mathbb{E}[Z] = 0$
- On the other hand by integration by parts

$$\begin{aligned}
\mathbb{E}[Z^2] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2 e^{-x^2/2} dx \\
&= -\frac{1}{\sqrt{2\pi}} \left(\left[x e^{-x^2/2} \right]_{-\infty}^{+\infty} - \int_{-\infty}^{+\infty} e^{-x^2/2} dx \right) \\
&= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1
\end{aligned}$$

Gaussian Distribution

Exercise

Let $Z \sim \mathcal{N}(0,1)$ and let $X = \sigma Z + \mu$ for $\sigma > 0, \mu \in \mathbb{R}$

1. Compute $\mathbb{E}[X]$, $\mathrm{Var}(X)$
2. Compute the p.d.f. of X

Solution

- By the properties of the expectation and the variance, $\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$
- On the other hand

$$F_X(t) = \mathbb{P}(X \leq t) = \mathbb{P}(\sigma Z + \mu \leq t) = \mathbb{P}\left(Z \leq \frac{t-\mu}{\sigma}\right) = \Phi\left(\frac{t-\mu}{\sigma}\right)$$

- Therefore

$$f_X(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{(x-\mu)^2/2\sigma^2}$$

Generic Gaussian Distribution

Motivation

- From the standard normal distribution we can define a whole family of normal distributions as $X = \sigma Z + \mu$
- These distributions are entirely characterized by their mean and their variance

Definition

*Let $\mu \in \mathbb{R}$ and $\sigma > 0$, a r.v. X has **the normal/Gaussian distribution with mean μ and variance σ^2** if X has th p.d.f.*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/(2\sigma^2)} \quad \text{for } x \in \mathbb{R}$$

We denote it $X \sim \mathcal{N}(\mu, \sigma^2)$

Generic Gaussian distribution

Exercise

Let $\mu \in \mathbb{R}$, $\sigma > 0$ and $X \sim \mathcal{N}(\mu, \sigma^2)$

Let $a \neq 0$ and $b \in \mathbb{R}$, show that $Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

In particular what is the dist. of $Z = \frac{X-\mu}{\sigma}$?

Solution

- Consider $a > 0$ (same can be done for $a < 0$)

$$\mathbb{P}(aX + b \leq t) \leq \mathbb{P}(X \leq \frac{t-b}{a}) = F_X(\frac{t-b}{a})$$

- so

$$f_Y(y) = \frac{1}{a} f_X(\frac{t-b}{a}) = \frac{1}{\sqrt{2\pi\sigma^2 a^2}} \exp(-\frac{(x-\mu a-b)^2}{2\sigma^2 a^2})$$

- In particular $Z \sim \mathcal{N}(0, 1)$

Generic Gaussian distribution

From generic to standard normal

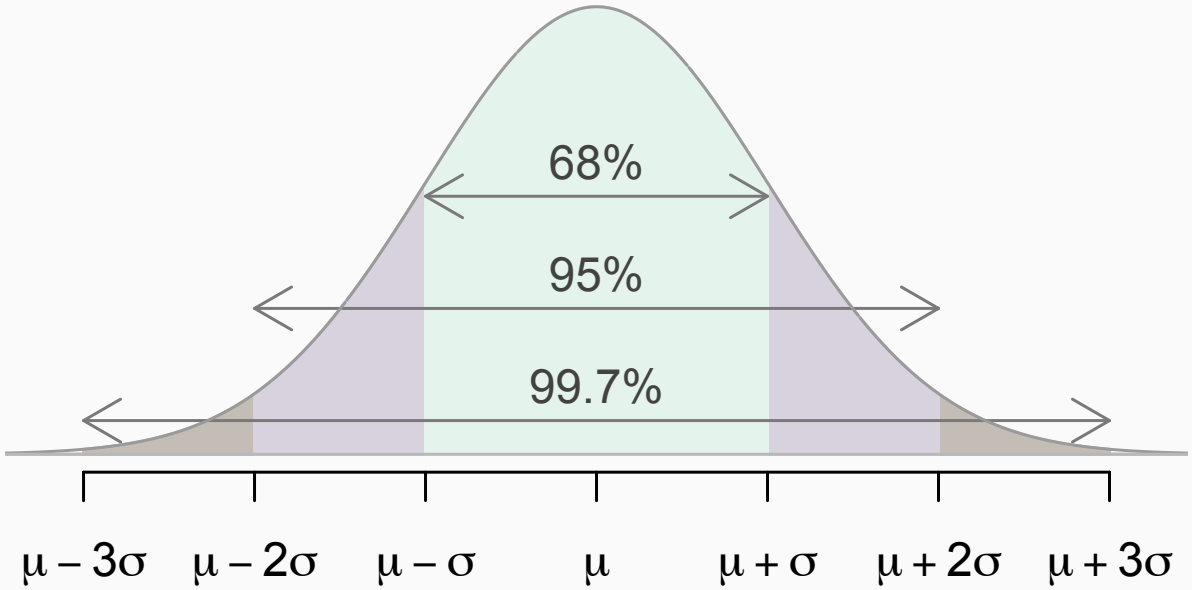
- Computing prob. of $X \sim \mathcal{N}(\mu, \sigma^2)$ can be done by using the c.d.f. of the standard normal dist.

$$\mathbb{P}(X \in [a, b]) = \mathbb{P}(a \leq X \leq b) = \mathbb{P}\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right),$$

$$\Rightarrow \quad \boxed{\mathbb{P}(X \in [a, b]) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)}$$

68-95-99.7 Rule

- X has a normal distribution,
  - about 68% probability X falls within 1 SD of the mean,
  - about 95% probability X falls within 2 SD of the mean,
  - about 99.7% probability X falls within 3 SD of the mean.
- The probability of X falls 4, 5, or more standard deviations away from the mean is very low.



Outline

Normal Approximation

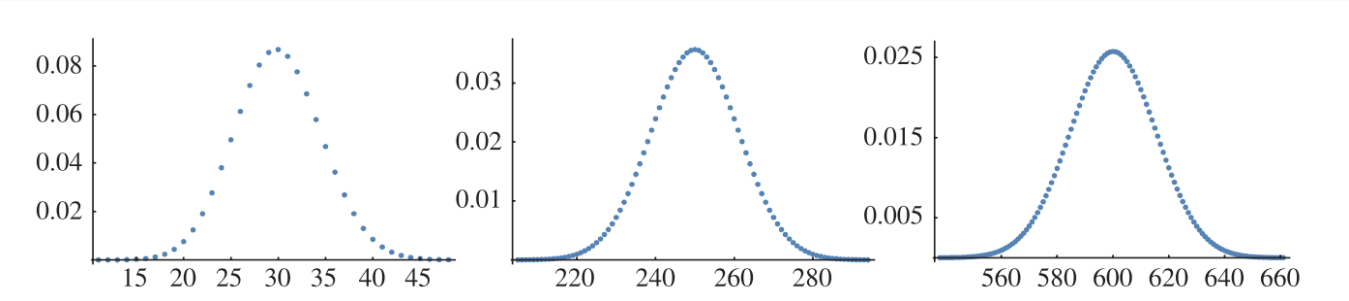
Motivation

- We turn back to our original motivation:
  How close is the empirical mean to the true mean of a r.v.?
- Here we focus on a flip of a coin ($X \sim \text{Ber}(p)$)
- Analyzing the empirical mean amounts to analyze the binomial

$$S_n = X_1 + \ldots + X_n \quad \text{for } X_i \overset{i.i.d.}{\sim} \text{Ber}(p)$$

  as $n \to +\infty$
- So we want to understand how a binomial looks like as n increases for a given fixed p

Normal Approximation

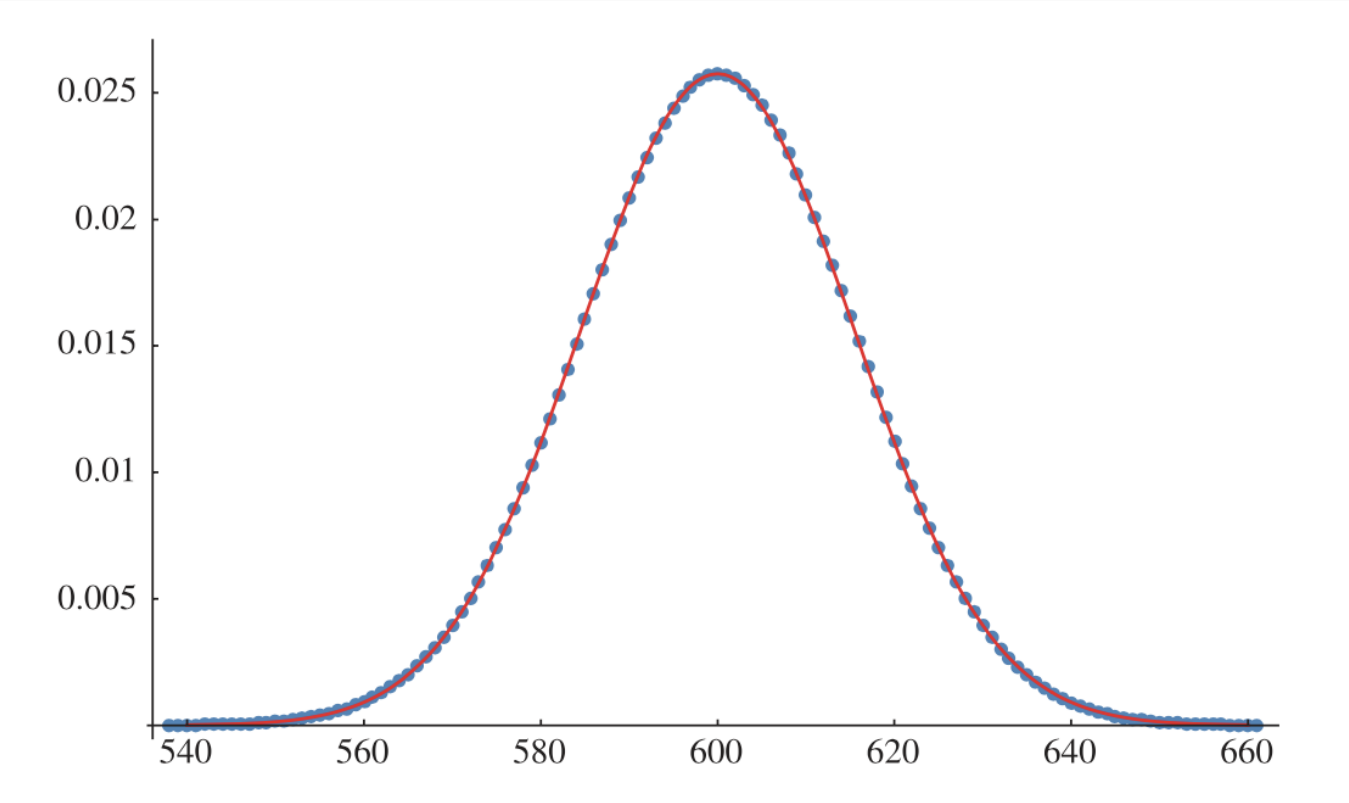


Plots of $S_{100} \sim \text{Bin}(100, 0.4)$, $S_{500} \sim \text{Bin}(500, 0.5)$, $S_{1000} \sim \text{Bin}(1000, 0.6)$

$\rightarrow$ Looks like the bell of a Gaussian distribution!

Figure from Introduction to probability, D. Anderson, T. Seppäläinen, B. Valkò

Normal Approximation



Bullets: $S_{1000} \sim \text{Bin}(1000, 0.6)$

Red curve: $X \sim \mathcal{N}(600, 240)$

Figure from Introduction to probability, D. Anderson, T. Seppäläinen, B. Valkò

Normal Approximation

Formalization

- If $X \sim \mathcal{N}(\mu, \sigma^2)$ approximates $S_n \sim \text{Bin}(n, p)$, we should have

$$\mathbb{E}[X] = \mathbb{E}[S_n] \quad \text{Var}(X) = \text{Var}(S_n)$$

- So $\mu = np$, $\sigma^2 = np(1-p)$, i.e. $X \sim \mathcal{N}(np, np(1-p))$
- Rather than comparing S_n to any r.v., let us *standardize* S_n to compare it to a standard normal r.v.

Standardization

- For a given r.v. Y, standardizing Y amounts to consider

$$\tilde{Y} = \frac{Y - \mu}{\sigma} \quad \text{for } \mu = \mathbb{E}[Y],\ \sigma^2 = \text{Var}(Y)$$

  such that $\mathbb{E}[\tilde{Y}] = 0$, and $\text{Var}(\tilde{Y}) = 1$
- For example, if $X \sim \mathcal{N}(\mu, \sigma^2)$,
  then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$ is its standardization

Normal approximation

Idea

- After standardization, we should have that

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}} = \frac{S_n - np}{\sqrt{np(1-p)}} \approx \frac{X - \mu}{\sigma} = Z \sim \mathcal{N}(0,1)$$

Theorem (Central limit theorem for binomial random variables)
Let $0 < p < 1$, *assume that* $S_n \sim \mathrm{Bin}(n,p)$.

Then for any $-\infty \leq a \leq b \leq +\infty$,

$$\lim_{n \to +\infty} \mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b \right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

Notes:

- Compared to the law of large numbers, this is a limit in distribution, i.e., as $n \to +\infty$, we get a formulation of the prob. in terms of a fixed p.d.f.

Normal approximation

Application

- Previous theorem is still only valid for a limit, below is a practical rule of thumb

Lemma
Suppose that $S_n \sim \text{Bin}(n, p)$ *with* n *large and* p *not too close to 0 and 1, then*

$$\mathbb{P}\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) \approx \Phi(b) - \Phi(a)$$

with Φ *the c.d.f. of* $Z \sim \mathcal{N}(0, 1)$.

As a rule of thumb the approx. is good if $np(1-p) > 10$.

Note:

- We will see that if p is too small even for large n the normal distribution is not the right approximation of the binomial.
- See backup-slides for finer approximations

Normal approximation

Practice

We roll a pair of fair dice 10,000 times.

Estimate the prob. that the number of times we get snake eyes (two ones) is between 280 and 300

Hint: Use the Central Limit Theorem and a table of the values of the c.d.f. of a standard normal dist.

Outline

Continuity correction

Continuity correction

- If $S_n \sim \text{Bin}(n, p)$, then it can only take integer values.
  Thus if k_1, k_2 are integers,

$$\mathbb{P}(k_1 \leq S_n \leq k_2) = \mathbb{P}(k_1 - 1/2 \leq S_n \leq k_2 + 1/2)$$

- The second interval is better to approximate the binomial, we can approx.

$$\mathbb{P}(k_1 - 1/2 \leq S_n \leq k_2 + 1/2) = \Phi\left(\frac{k_2 + 1/2 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k_1 - 1/2 - np}{\sqrt{np(1-p)}}\right)$$

- Typically if $k_1 = k_2$, the approx. of $\mathbb{P}(k_1 \leq S_n \leq k_2)$ by a normal dist. would ive 0 which is completely wrong
- The correction given above remedies this problem