



BITTIGER

DS 501 Data scientist express bootcamp

[Ella]



版权声明

所有太阁官方网站以及在第三方平台课程中所产生的课程内容，如文本，图形，徽标，按钮图标，图像，音频剪辑，视频剪辑，直播流，数字下载，数据编辑和软件均属于太阁所有并受版权法保护。

对于任何尝试散播或转售BitTiger的所属资料的行为，太阁将采取适当的法律行动。



有关详情，请参阅

<https://www.bittiger.io/termsfuse> <https://www.bittiger.io/termservice>

Copyright Policy

All content included on the Site or third-party platforms as part of the class, such as text, graphics, logos, button icons, images, audio clips, video clips, live streams, digital downloads, data compilations, and software, is the property of BitTiger or its content suppliers and protected by copyright laws.

Any attempt to redistribute or resell BitTiger content will result in the appropriate legal action being taken.



We thank you in advance for respecting our copyrighted content.

For more info:

see <https://www.bittiger.io/termsofuse>
and <https://www.bittiger.io/termsofservice>



What is statistics? Why is it important?



Statistics

- All about quantifying uncertainty (which subject is studying certainty?)
- Why uncertainty? **imperfect** data
 - Psychology
 - Geology
 - Economics
- Why imperfect data?
 - Only observe a small fraction, e.g., candidate poll
 - Only observe indirect signs, e.g., neuro signal
 - Data always contain noise, e.g., measurement error



Probability

- Probability is central to statistics
 - Probability = Statistics?
- Basic concepts
 - **Experiment**
 - **Sample space**: all possible outcomes of an **experiment**
 - **Event**: subset of sample space



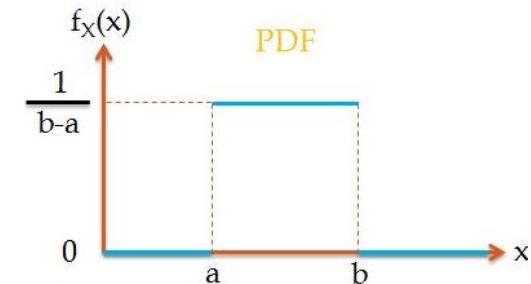
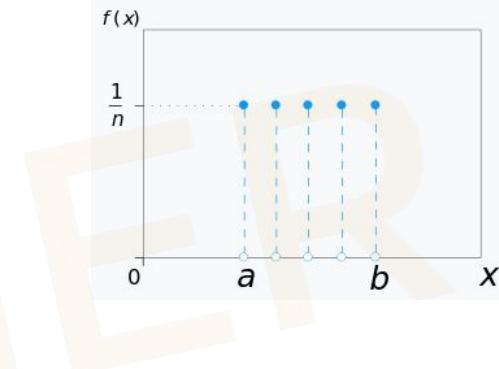
Random variable and types

- Random variable (X)
 - Enumerate sample space to real value
- Categories
 - Discrete
 - Continuous
- Examples
 - Flip a coin; Daily traffic; Number of people click on an ad; BMI
- R
 - bar chart, pie chart
 - Probability density



How to describe probability of random variable

- Probability density function (PDF)
 - PDF: $x \sim P(X = x)$
 - For X , frequency of possible outcome x 's occurrences of the total possible occurrences
 - Example: rolling dice, $P(X = 1) = \frac{1}{6}$
 - Discrete (pmf) v.s continuous (pdf)
 - Non-negative, sum up to 1
- Rules
 - $0 \leq P(X = x) \leq 1$, $P(X = \Omega) = 1$, $P(X = \emptyset) = 0$, $P(X = x) + P(X \neq x) = 1$, $P(X = x \text{ or } Y = y) = P(X = x) + P(Y = y) - P(X = x \text{ and } Y = y)$





Example interview questions

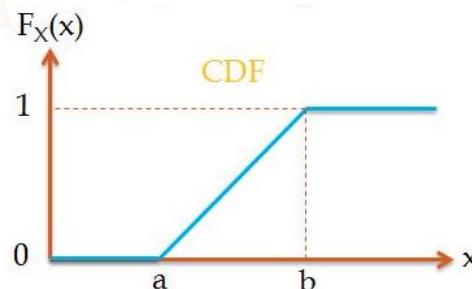
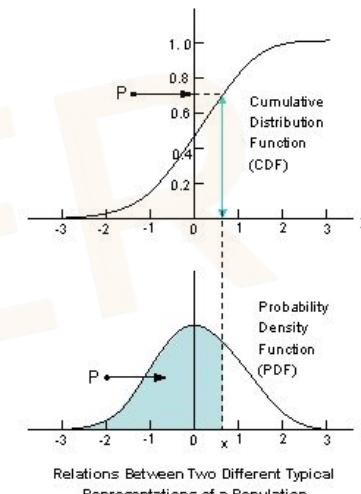
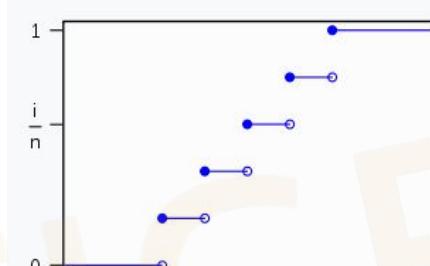
- 2 dices, probability of at least one 4 shows up
- 2 dices, probability of the sum is 6
- ...

BITTIGER



How to describe probability of random variable(cont'd)

- Cumulative density function (CDF)
 - CDF: $x \sim P(X \leq x)$
 - Discrete v.s continuous
 - Non-decreasing, up to 1
- Examples
 - Roll dice
 - Uniform distribution

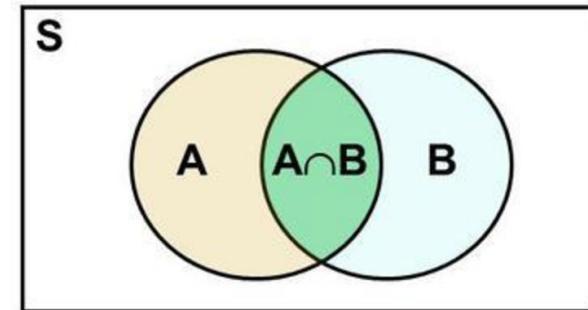




Conditional probability

- How to update uncertainty after getting new information?
 - $P(\text{dice} = 1) = \frac{1}{6}$, what if we know it's an odd number?
- Definition, Probability of A given B
- Intuition
 - Only consider B
 - Renormalize
- Example question

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$





Independence

- Def: $P(A \cap B) = P(A) \times P(B)$
- Considering conditional probability $P(A | B) = \frac{P(A \cap B)}{P(B)}$

$$P(A | B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

- A and B are conditionally independent given C

$$P(AB | C) = P(A|C)P(B|C)$$

- Example question



A and B are independent \leftrightarrow
A and B are independent given C ?



Bayes' theorem

- Reverse the conditional probability

$$P(B | A) = \frac{P(A | B)P(B)}{P(A | B)P(B) + P(A | B^c)P(B^c)}$$

- Prove it
 - Hint:
- Fair coin and unfair coin problem

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$



Expectation

- Measure what's the center of a random variable
 - Discrete $E[X] = \sum_x xp(x)$
 - Continuous $E(X) = \int_x x P(X=x)$
 - Roll dice, flip coin, uniform distribution
- Population mean v.s. sample mean
 - Sample mean **itself** is random variable, estimator of population mean
 - Unbiased: $E(\text{sample mean}) = \text{population mean}$
- Why always larger sample is better?
 - Biased when fewer data?
 - Proof. Hint: $E(X + Y) = E(X) + E(Y)$



Variance

- Variance (σ^2) measure how random variable spreads

$$Var(X) = E[(X - \mu)^2] = E[X^2] - E[X]^2$$

- Calculate variance from rolling dice

- Standard deviation (σ)

- Sample variance (s^2) and standard deviation (s)

$$S^2 = \frac{\sum_{i=1} (X_i - \bar{X})^2}{n - 1}$$

- why $(n-1)$: unbiased estimator of σ^2
 - Proof (HW), Hint $Var(X+Y) = Var(X) + Var(Y)$ if X, Y independent



Important discrete distribution

- Bernoulli

- $P(X = x) = p^x(1 - p)^{1-x}$

- Binomial

- Sum of N Bernoulli variables
 -

$$P(X = x) \equiv \binom{N}{x} p^x(1 - p)^{N-x}$$

$$E(X) = p, \text{Var}(X) = p(1-p)$$

What's the mean and variance?

$$E(X) = N*p, \text{Var}(X) = N*p(1-p)$$



More details about other distributions

- Geometric distribution

- Prob of seeing first success with k independent trials, each with success probability p.

$$\Pr(X = k) = (1 - p)^{k-1} p$$

- Negative geometric distribution

- Prob of seeing r failures with k success trials , each with success probability p.

$$\Pr(X = k) = \binom{k + r - 1}{k} p^k (1 - p)^r$$



Important distribution

- Poisson, count data per time unit

- $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$

- Examples

- Exponential distribution

- In a poisson distribution, inter-arrival time follows exponential distribution

- PDF: $\lambda e^{-\lambda x}$ CDF: $1 - e^{-\lambda x}$

- Expectation

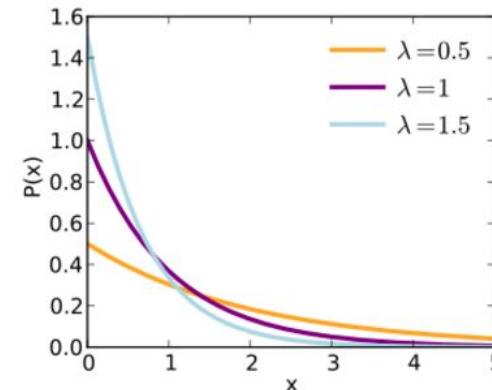
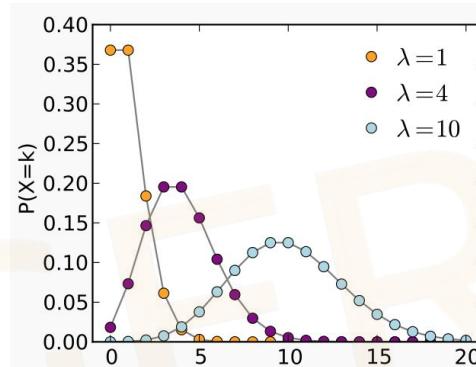
$$E[X] = \frac{1}{\lambda}$$

- Variance

$$\text{Var}[X] = \frac{1}{\lambda^2}$$

- “Memoryless”

$$P(X > s+t \mid X > s) = P(X > t)$$





Important continuous distribution

- Normal distribution

$$P(X = x) = (2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}$$

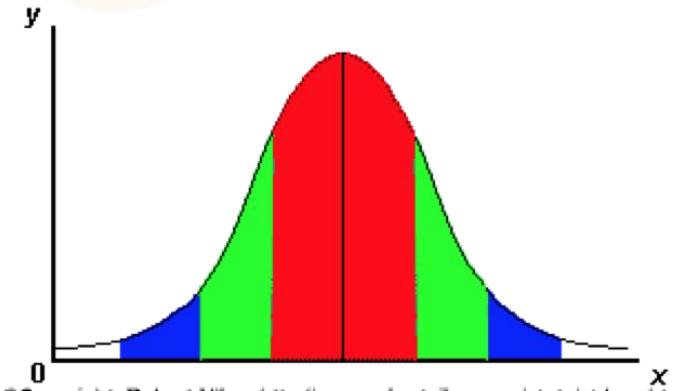
$$E[X] = \mu \text{ and } Var(X) = \sigma^2$$

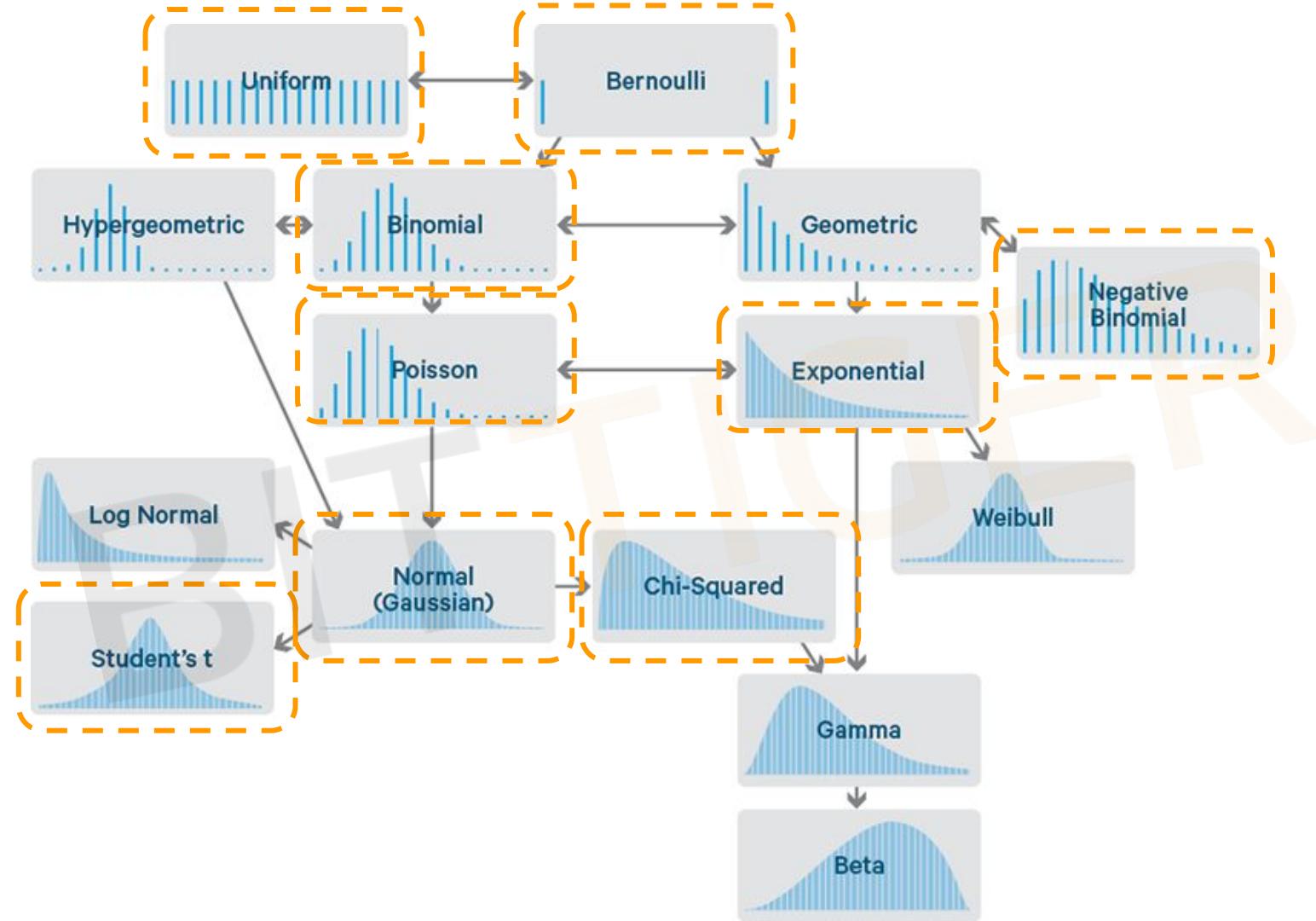
- Standard normal distribution (Z) $\mu = 0$ and $\sigma = 1$

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

- Beauty of normal curve (6σ)

- [$\mu - 3\sigma, \mu + 3\sigma$] covers 99.7%
- [$\mu - 2\sigma, \mu + 2\sigma$] covers 95%
- [$\mu - \sigma, \mu + \sigma$] covers 68%







Example interview questions

- In a country in which people only want boys every family continues to have children until they have a boy. If they have a girl, they have another child. If they have a boy, they stop. What is the proportion of boys to girls in the country?
 - What distribution is this?
 - What's expected number of kids (X) each family has?
 - What the ratio of boys to girls?



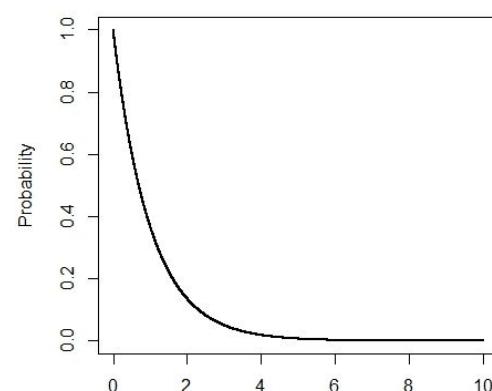
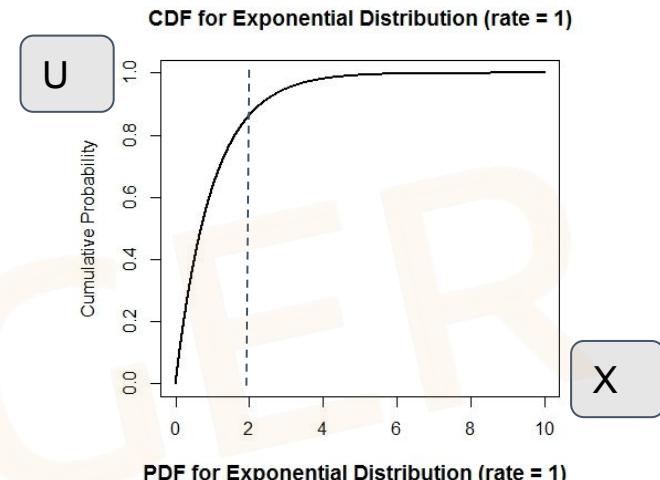
Another type of distribution Interview question

- How can we generate certain distribution?
 - Q1: Generate exponential distribution from uniform distribution
 - Q2: Generate normal distribution from uniform distribution



Inverse transform sampling

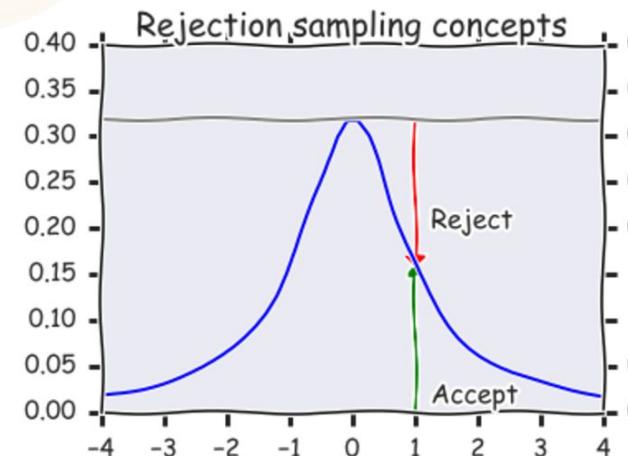
- To generate X with PDF: $P(X)$
 - Calculate CDF of X : $F(X)$ and inverse $F^{-1}(\cdot)$
 - Generate U with uniform distribution, say u
 - Calculate $F^{-1}(u)$
 - Repeat n times
- Intuition
- Let's solve Q1
- Can we solve Q2 as well?





Acceptance rejection sampling

- Simple version: if good, accept; else, reject.
 - Ex1: Generate Y with uniform $[1/4, 1]$ given $X \sim [0, 1]$
 - Ex2: Generate random integer Y in $[1, 7]$ given random integer X in $[1, 5]$ $5*(X-1) + X$
 - Ex3: Generate normal distribution given uniform distribution
 - Sample a point, x , calculate $P(x)$
 - Generate random number u from uniform distribution
 - If $u < P(x)$, accept x , else reject, back to step 1
- Check out this [video](#)





Percentile

- Definition ($c\%$, N)



- Special percentiles
 - Median
 - 1st quantile
 - 3rd quantile
- Median, quantiles in normal distribution?
 - Look up z score, alpha: area outside $[-z, z]$



Why normal distribution is so important?

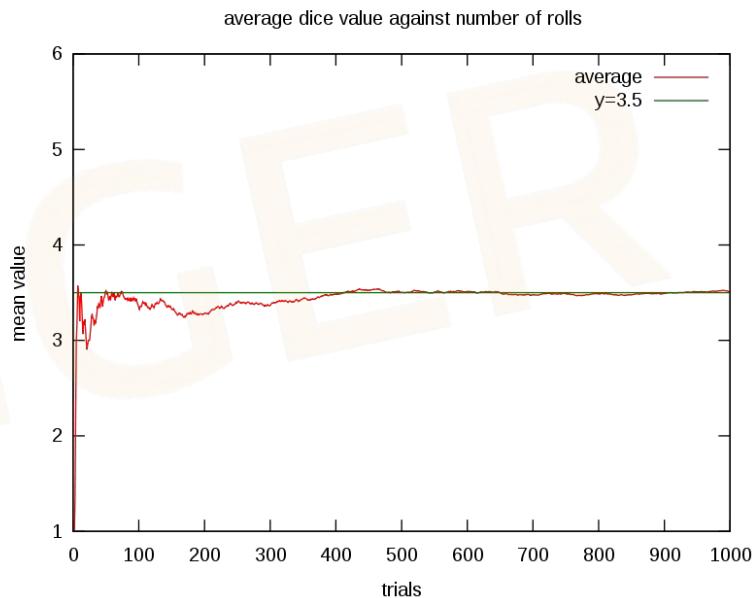


Law of large numbers

$X_1, X_2 \dots X_n \dots$ are independent, identically - distributed (IID) random variables, X has finite mean μ .

- Sample mean $\bar{X}_n \equiv \frac{1}{n} \sum_{i=1}^n X_i$ converge to the true mean μ as n increases:

$$\bar{X}_n \xrightarrow{n \rightarrow \infty} \mu$$





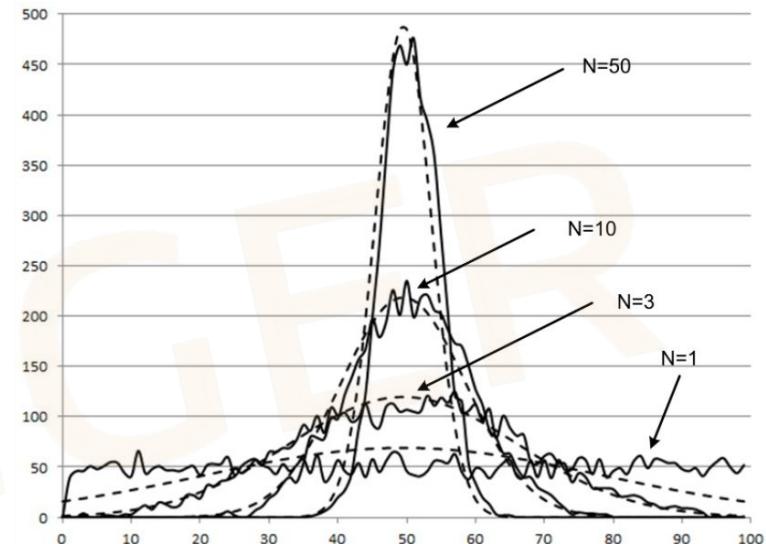
Central limit theorem

- $X_1, X_2, \dots, X_n \dots$ are independent, identically-distributed (IID) random variables, X has finite mean μ and variance σ^2

$$\bar{X}_n \equiv \frac{1}{n} \sum_{i=1}^n X_i \sim N(\mu, \sigma^2/n)$$

(replacing σ by sample standard error, CLT still holds)

- Application
 - Binomial distribution
 - Flip coin example



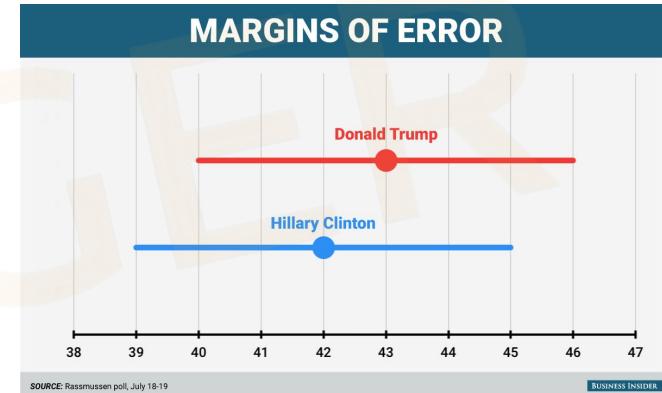


Confidence intervals

- Recall normal curve
 - What range covers 95% of possibility? 99%?
- Confidence interval (CI)
 - Derive 95% CI of p in Binomial distribution

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

- How many voters in the candidate poll?
- Can Trump relax? How many are enough?





When CLT doesn't hold?

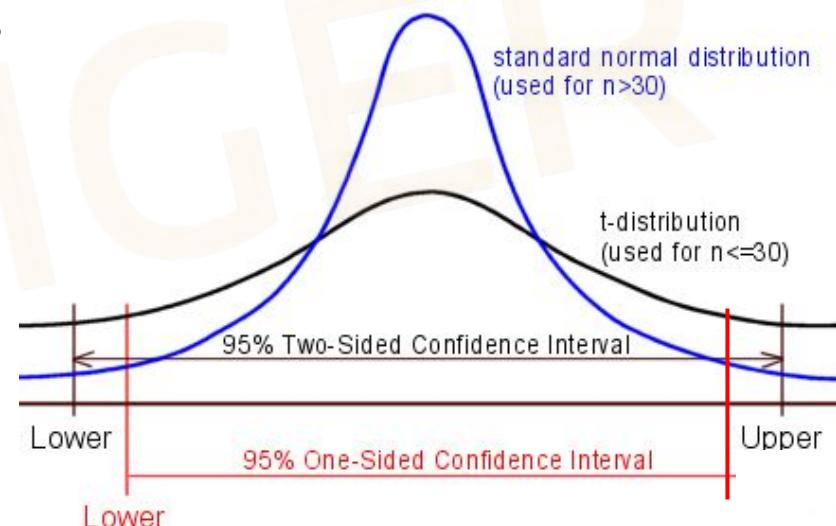
- Normal distribution -> T distribution when $N < 30$
 - T distribution has only one parameter: degree of freedom ($df = N-1$)
 - Approximate normal as df increases
 - CI under normal distribution

$$Mean_{estimate} \pm z_{1-\alpha/2} * StdErr_{estimate}$$

- CI under t distribution

$$Mean_{estimate} \pm t_{n-1} * StdErr_{estimate}$$

- z or t?



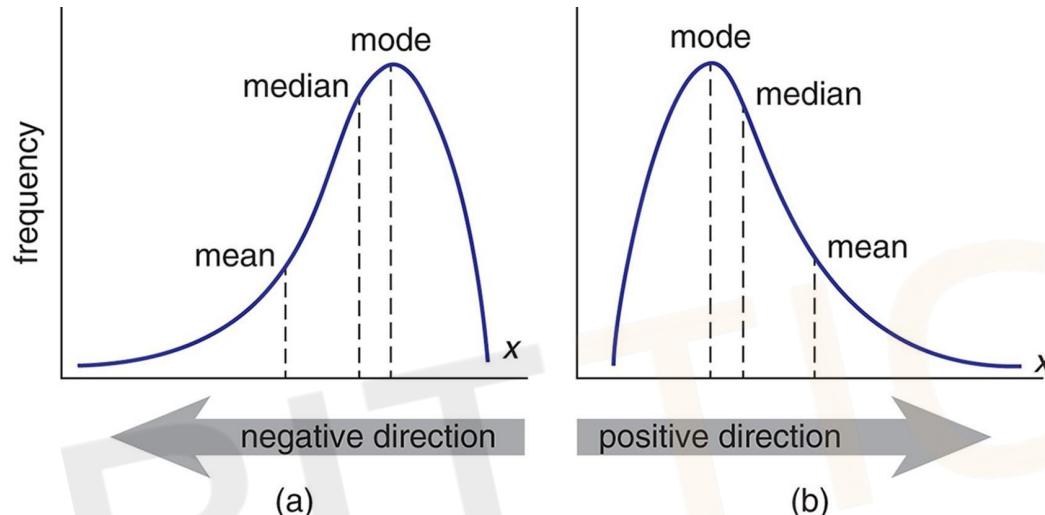


BIT TIGER

What about CI of not normal distribution?



Skewed distribution



- What caused skewness
 - Bounded: lower bound or upper bound
- Can we still use normal distribution to estimate mean of skewed variable?



How to treat skewed variable

- Log transformation
- Use median, if so, what's CI of median?
 - Non-parametric method (v.s. parametric method)
 - Bootstrap
 - Jackknife



Covariance and correlation

- Definition

- $cov(X, Y) = E[(X - E[X])(Y - E[Y])]$

- $\rho_{X,Y} = corr(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$

- What's sample covariance and correlation?

- $cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1} (\sum x_i y_i - n \bar{x} \bar{y})$

- $cor(X, Y) = \frac{cov(X, Y)}{s_x s_y}$



Properties of correlation

- Properties

- $\text{cor}(X, Y) = \text{cor}(Y, X)$
- $-1 \leq \text{cor}(X, Y) \leq 1$
- $\text{cor}(X, Y) = 1$ or -1 : X, Y aligned on a line perfectly
- $\text{cor}(X, Y) = 0$: X, Y are independent

- Geometric interpretation

- $$\rho_{X, Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} = \frac{\sum_{i=1}^n x'_i y'_i}{\|x'\| \|y'\|} = \cos(x', y')$$
- $\text{cor}(X, Y) = 0, \text{cor}(Y, Z) = 0 \Rightarrow \text{cor}(X, Z) = 0?$

