# Central Limit Theorem

Eli Gurarie

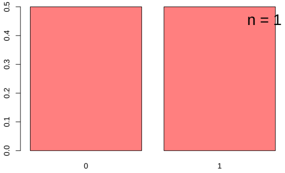
StatR 101 - Lecture 8a
November 15, 2012

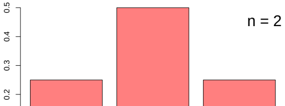
November 15, 2012

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

## Coinflips

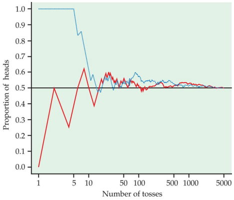
What is the distribution of a Bernoulli trial, repeated many, many times?



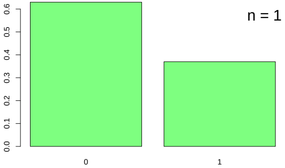
**Law of Large Numbers**

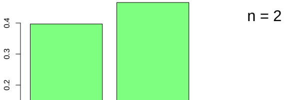
If you repeat an experiment $X$ many, many, many times ($i = (1, 2, 3, ...., n)$, the average of $X$ will asymptotically aproach $E(X)$.




## Shaq shoots

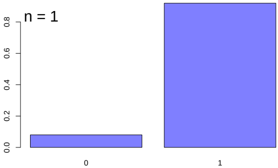
What about an assymetric distribution ($p = 0.37$)?

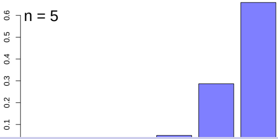


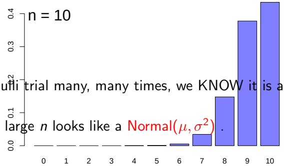
## Ray Allen shoots

Fine! What about an *extremely* assymetric distribution ($p = 0.92$)?



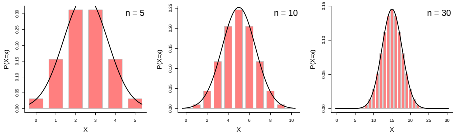

## A hypothesis:



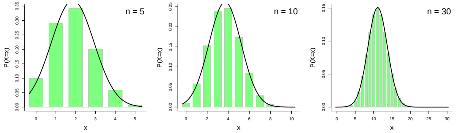

- If we repeat an Bernoulli trial many, many times, we KNOW it is a binomial...
- But Binomial($n, p$) at large $n$ looks like a Normal($\mu, \sigma^2$) .

# But what are the mean and variance?

- Match the binomial distribution's mean ($np$) and variance ($np(1-p)$)




$$p = 0.5,\ n = (5, 10, 30),$$
$$\mu = 2.5, 5, 15,$$
$$\sigma^2 = 1.25, 2.54, 7.5$$




**Normal approximation to the Binomial**

- A variable $X \sim \text{Binomial}(n, p)$at large $n$ is approximated by a continuous normal distribution

$$\mathcal{N}(\mu = np, \sigma^2 = np(1-p))$$

- This is useful because: $n!$ can be difficult to compute.

$$\mu = 4.6, 9.2, 27.6,$$

**Caution:**

The normal distribution is *continuous* - so it can not (easily) tell you the probability of a single discrete value ($P(X = x)$) ... but it is quite good for calculating ranges ($P(a < X < b)$).

# Example of normal approximation of binomial

About 4% of students have tattoos. Let $X$ be the number of students that have tattoos in a review section of $n_1 = 30$, and $Y$ be the number of students that have tattoos in a lecture of $n_2 = 200$ students. **What is the probability that no more than 3 students have a tattoo? ($Pr(X \leq 3)$ and $Pr(Y \leq 3)$)**

- $X \sim \text{Binomial}(n_1 = 30, p = 0.04)$, is approximated as: $X \sim \mathcal{N}(\mu = 1.2, \sigma^2 = 1.152)$
- $Y \sim \text{Binomial}(n_2 = 200, p = 0.04)$, is approximated as: $Y \sim \mathcal{N}(\mu = 8, \sigma = 2.77)$

True values:

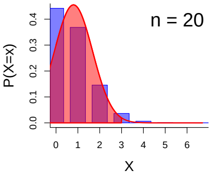
- $P(X \leq 3) = \sum_{i=0}^{3} f(x|30, .04) =$ `pbinom(3,n1,sqrt(n1*p*(1-p)))` $= 0.9694$
- $P(Y \leq 3) = \sum_{i=0}^{3} f(y|200, .04) =$ `pbinom(3,n2,sqrt(n2*p*(1-p)))` $= 0.0395$

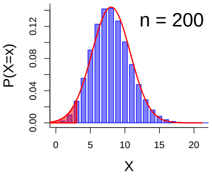
Approximate values:

- $P(X \leq 3) = \int_{-\infty}^{3} f(x|1.2, 1.152) =$ `pnorm(3,mean=1.2, sd=1.07)` $= 0.9532$
- $P(Y \leq 3) = \int_{-\infty}^{3} f(y|8, 7.68) =$ `pnorm(3,mean=8, sd=2.77)` $= 0.0355$

# Example of normal approximation of binomial

Note that the approximation is best for higher $n$, but the estimates are worse away from the mean of the distribution.




- True: $P(X \leq 3) = 0.9694$
- Approx: $P(X \leq 3) = 0.9532$

- True: $P(Y \leq 3) = 0.0355$
- Approx: $P(Y \leq 3) = 0.0395$

# A hypothesis:

- If we repeat an Bernoulli trial many, many times, it looks like a Normal$(\mu, \sigma)$ distribution.

# But what are the mean and variance?

- Recall the summation rules of Expectation and Variance
  - $\text{E}(X_1 + X_2 + X_3 + ...) = \text{E}(X_1) + \text{E}(X_2) + \text{E}(X_3) + ...$
  - More generally
$$\text{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{E}(X_i)$$
  - $\text{Var}(X_1 + X_2 + X_3 + ...) = \text{Var}(X_1) + \text{Var}(X_2) + \text{Var}(X_3) + ...$
  - More generally
$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i)$$

  (**Note:** the variance rule is only for independent $X$.)

**Normal approximation to many Bernoulli trials**

- If a variable $Y = \sum_{i=1}^{n} X_i$ where $X_i$ is a Bernoulli random variable with probability $p$ $(X \sim \text{Bernoulli}(p))$ , then (when $n$ is large), $Y$ is distributed approximately:

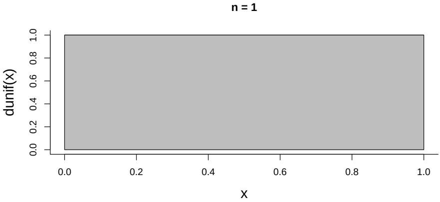
$$\mathcal{N}(\mu = np, \sigma = \sqrt{np(1-p)})$$

- Recall that $\mathrm{E}\,(X) = p$ and $\mathrm{Var}\,(X) = p(1-p)$ ... so we can say (in this case) that:

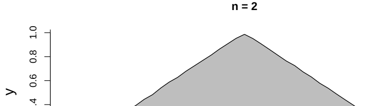
$$\mathcal{N}(\mu = n\,\mathrm{E}\,(X)\,, \sigma = \sqrt{n\,\mathrm{Var}\,(X)})$$

## What about other distributions?

$X \sim \text{Unif}(0,1)$

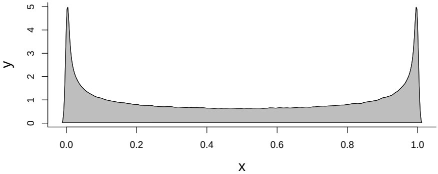



$Y = X_1 + X_2$

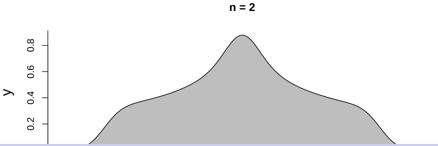


## What about a crazy distributions?

$X \sim \text{Beta}(.5, .5)$ .... $\text{E}(X) = 1/2$, $\text{Var}(X) = 1/8$




$Y = X_1 + X_2$




**Central Limit Theorem (CLT)**

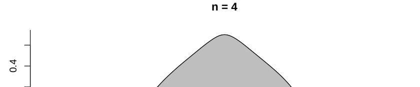
If $X_1, X_2, X_3 \ldots X_n$ are **any, independent, identically distributed (iid)** random variables with mean $\mu_x$ and variance $\sigma_x^2$, and

$$Y = \sum_{i=1}^{n} X_i$$

then, as $n$ becomes large

$$Y \sim \mathcal{N}(n\mu_x, n\sigma_x^2)$$

**In words**: If you add up a BUNCH OF IID RANDOM VARIABLES, the result will be distributed approximately as a NORMAL distribution!



## Central Limit Theorem (CLT)

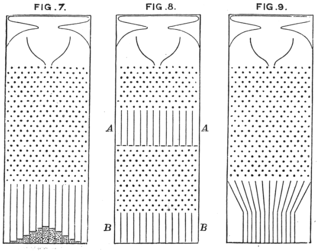
- The single most important theorem in Statistics!
- Because:
  - Many processes in nature are “additive” - (e.g. growth)
  - Many statistical objects (including many we have seen) are based on sums!
    - Remember: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
    - According to the CLT, if $X$ has a distribution with mean $\mu_x$ and $\sigma_x$,

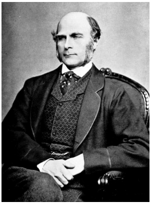
      then

$$\overline{X} \sim \mathcal{N}\left(\mu_x, \frac{\sigma_x^2}{n}\right)$$

  - This fact is the basis of A LOT of **classical inference** techniques!

## Central Limit Theorem: Galton’s Box (the Quincunx)



**Francis Galton** (1822-1911)
Founder of regression, correlation, weather maps, fingerprinting, questionnaires...

Video: http://www.youtube.com/watch?v=9xUBhhM4vbM