Chapter 4: Distributions of random variables

Wen-Han Hwang
(Slides primarily developed by Mine Çetinkaya-Rundel from OpenIntro.)

Institute of Statistics
National Tsing Hua University
Taiwan

OpenIntro

Outline

Normal distribution

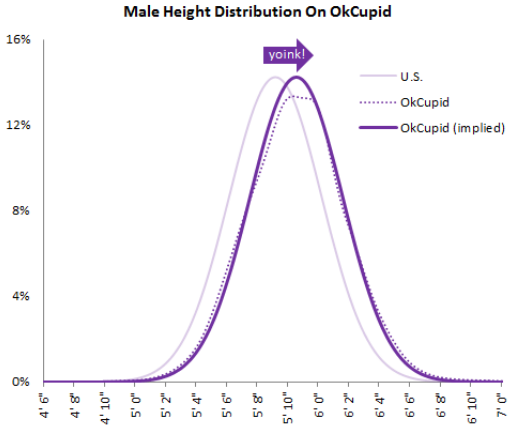
Normal distribution

- Unimodal and symmetric, bell shaped curve
- Many variables are nearly normal, but none are exactly normal
- Denoted as $N(\mu, \sigma^2) \rightarrow$ Normal with mean μ and standard deviation σ

We denote $N(\mu=2, \sigma=4)$ or $N(2, 4^2)$


$f(x)$
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
0,4
0,3
0,2
0,1
0
1
2
3
4
5
6
x
$\mu - \sigma$
μ
$\mu + \sigma$

Heights of males

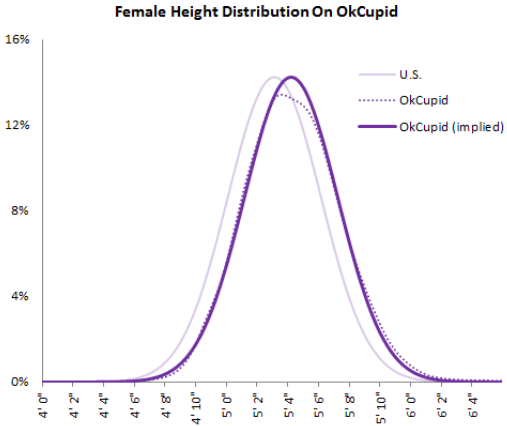


"The male heights on OkCupid very nearly follow the expected normal distribution – except the whole thing is shifted to the right of where it should be. Almost universally guys like to add a couple inches."

"You can also see a more subtle vanity at work: starting at roughly 5′ 8″, the top of the dotted curve tilts even further rightward. This means that guys as they get closer to six feet round up a bit more than usual, stretching for that coveted psychological benchmark."

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

Heights of females



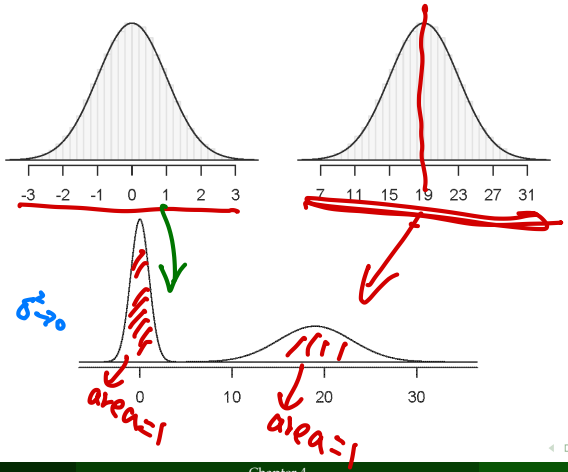
"When we looked into the data for women, we were surprised to see height exaggeration was just as widespread."

http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating/

Normal distributions with different parameters

μ: mean, σ: standard deviation

$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$



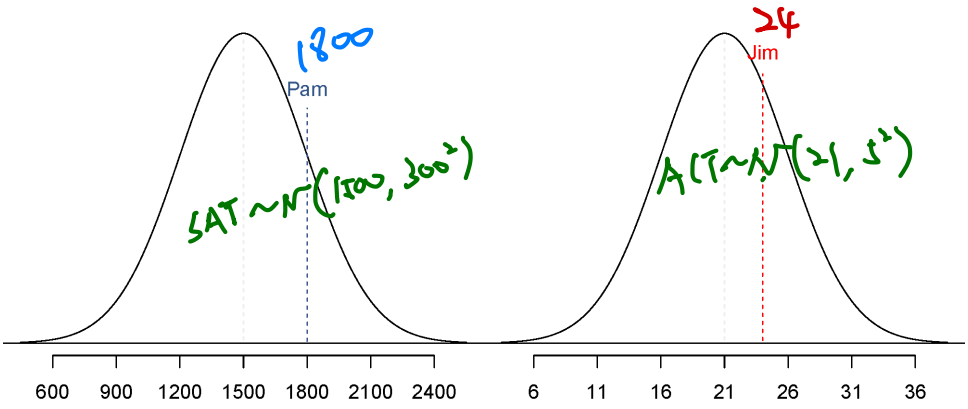
$N(\mu, \sigma^2) \Longrightarrow \delta_\mu$

$\sigma^2 \to 0$

$\downarrow$

$P(X=\mu)=1$

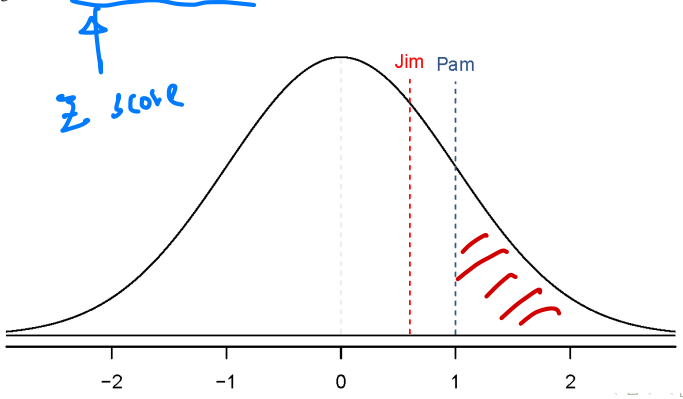
SAT scores are distributed nearly normally with mean 1500 and standard deviation 300. ACT scores are distributed nearly normally with mean 21 and standard deviation 5. A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers: Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?



Standardizing with Z scores

Since we cannot just compare these two raw scores, we instead compare how many standard deviations beyond the mean each observation is.

- Pam's score is $\frac{1800-1500}{300} = 1$ standard deviation above the mean.
- Jim's score is $\frac{24-21}{5} = 0.6$ standard deviations above the mean.



Standardizing with Z scores (cont.)

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

$\frac{X-\mu}{\sigma} \sim N(0,1)$

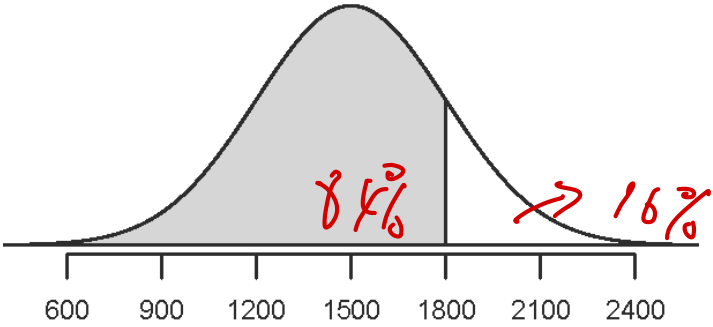
- These are called standardized scores, or Z scores.
- Z score of an observation is the number of standard deviations it falls above or below the mean.

$$Z = \frac{observation - mean}{SD}$$

- Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.
- Observations that are more than 2 SD away from the mean ($|Z| > 2$) are usually considered unusual.

Percentiles

- Percentile is the percentage of observations that fall below a given data point.
- Graphically, percentile is the area below the probability distribution curve to the left of that observation.



Calculating percentiles - using computation

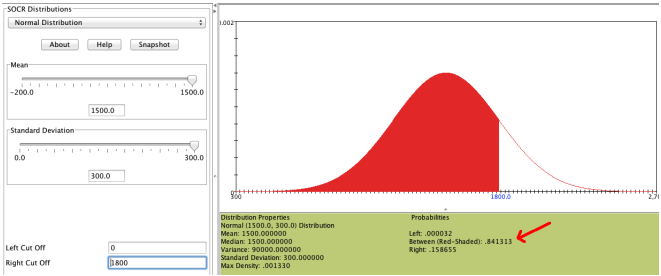
There are many ways to compute percentiles/areas under the curve:

- R:

```
> pnorm(1800, mean = 1500, sd = 300)
[1] 0.8413447
```

? pnorm

- Applet: `https://gallery.shinyapps.io/dist_calc/`



Calculating percentiles - using tables

	Second decimal place of Z									
Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

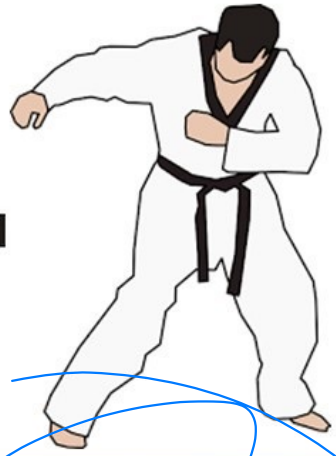
$P(N(0,1) \leq 0.67)$

Six sigma

"The term *six sigma process* comes from the notion that if one has six standard deviations between the process mean and the nearest specification limit, as shown in the graph, practically no items will fail to meet specifications."

http://en.wikipedia.org/wiki/Six_Sigma

What is Six Sigma



6σ

25K+
Views

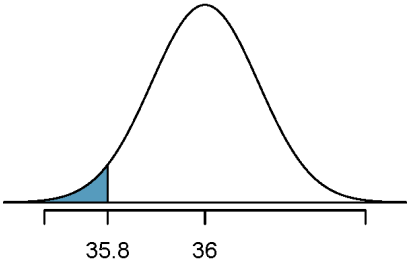
6σ = 99.9997%
3.4 Defects per Million Opportunities

Quality control

At Heinz ketchup factory the amounts which go into bottles of ketchup are supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz. Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection. What percent of bottles have less than 35.8 ounces of ketchup?

36 ± 0.2

Let X = amount of ketchup in a bottle: $X \sim N(\mu = 36, \sigma = 0.11)$



$$Z = \frac{35.8 - 36}{0.11} = -1.82$$

Finding the exact probability - using R

```
> pnorm(-1.82, mean = 0, sd = 1)
[1] 0.0344
```

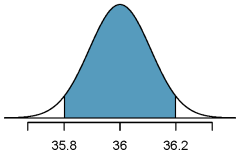
OR

```
> pnorm(35.8, mean = 36, sd = 0.11)
[1] 0.0345
```

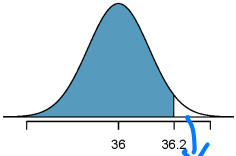
Practice

What percent of bottles pass the quality control inspection?

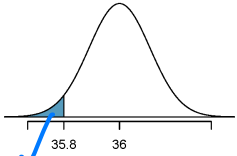
(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%



=



-



36 ± 0.2

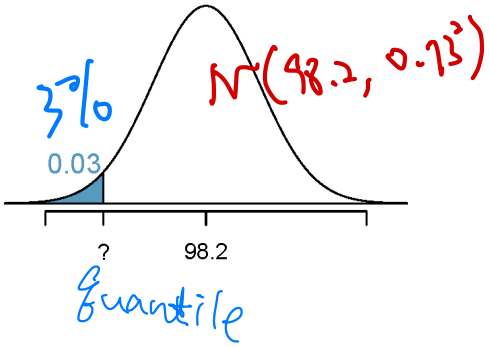
passed

0.0345

0.0345

Finding cutoff points

Body temperatures of healthy humans are distributed nearly normally with mean 98.2 F° and standard deviation 0.73F°. What is the cutoff for the lowest 3% of human body temperatures?



$$
\begin{aligned}
P(X < x) &= 0.03 \rightarrow P(Z < -1.88) = 0.03 \\
Z &= \frac{obs - mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = -1.88 \\
x &= (-1.88 \times 0.73) + 98.2 = 96.8F^\circ
\end{aligned}
$$

```
> qnorm(0.03)
[1] -1.880794
```

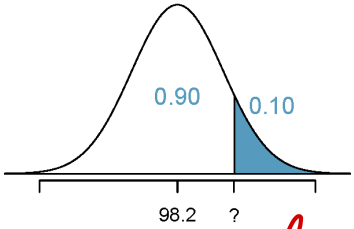
qnorm(0.03, 98.2, 0.73)

Mackowiak, Wasserman, and Levine (1992), *A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlick.*

Practice

Body temperatures of healthy humans are distributed nearly normally with mean $98.2F^\circ$ and standard deviation $0.73F^\circ$. What is the cutoff for the highest 10% of human body temperatures?

(a) $97.3\ F^\circ$
(b) $99.1\ F^\circ$
(c) $99.4\ F^\circ$
(d) $99.6\ F^\circ$



$$
\begin{aligned}
P(X > x) &= 0.10 \rightarrow P(Z < 1.28) = 0.90 \\
Z &= \frac{obs\ -\ mean}{SD} \rightarrow \frac{x - 98.2}{0.73} = 1.28 \\
x &= (1.28 \times 0.73) + 98.2 = 99.1
\end{aligned}
$$

qnorm(0.9)
or by Z-table

Calculating percentiles - using tables

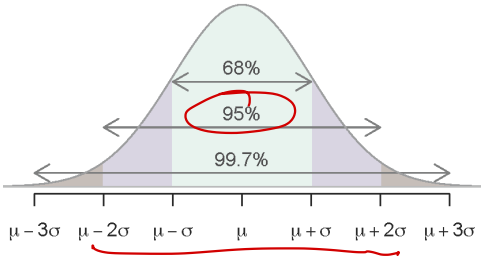
$\text{qnorm}(0.2) = -0.84$

Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

1.1 0.09

68-95-99.7 Rule

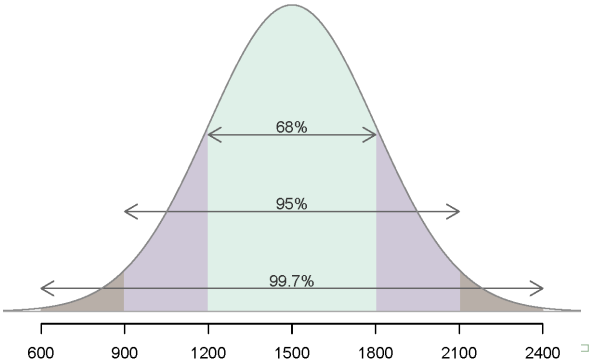
- For nearly normally distributed data,
  - about 68% falls within 1 SD of the mean,
  - about 95% falls within 2 SD of the mean,
  - about 99.7% falls within 3 SD of the mean.
- It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



Describing variability using the 68-95-99.7 Rule

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.

- ∼68% of students score between 1200 and 1800 on the SAT.
- ∼95% of students score between 900 and 2100 on the SAT.
- ∼99.7% of students score between 600 and 2400 on the SAT.



Practice

Which of the following is false?

(a) Majority of Z scores in a right skewed distribution are negative.

(b) In skewed distributions the Z score of the mean might be different than 0.

(c) For a normal distribution, IQR is less than $2 \times SD$. $P(|Z|<2) = 68\%$

(d) Z scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

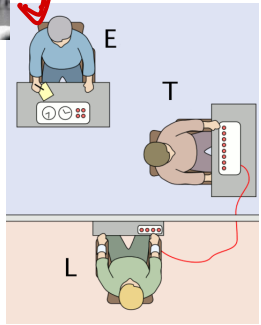
unusual data point: $|Z| > 2$.

very unusual $|Z| > 3$

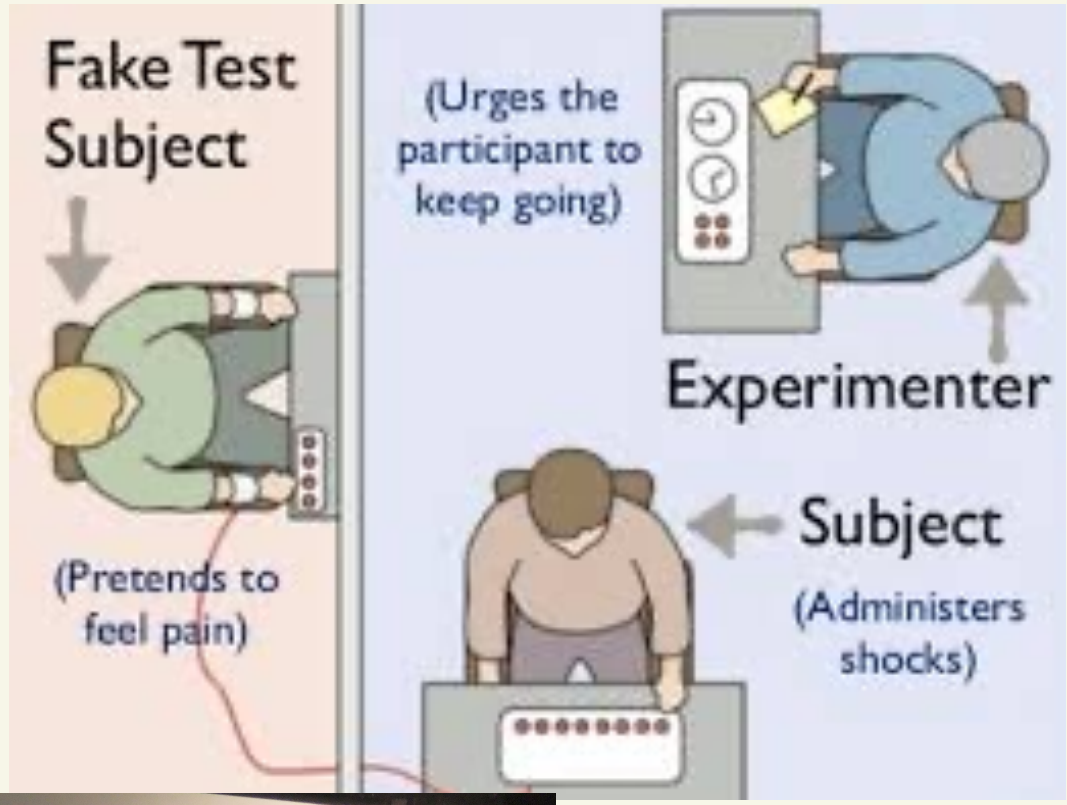
Geometric distribution

Milgram experiment

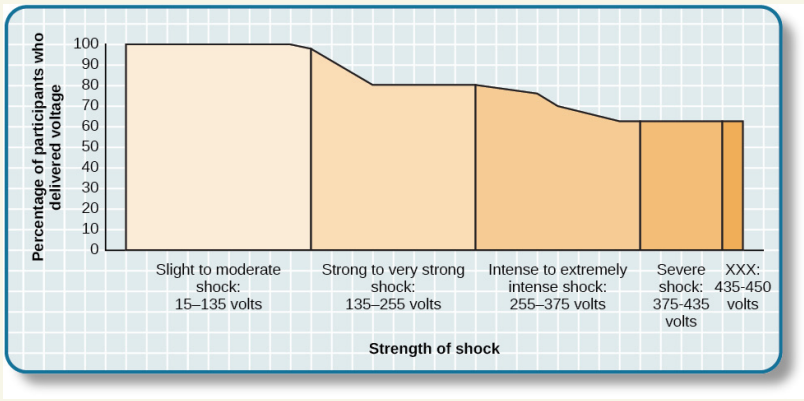
- Stanley Milgram, a Yale University psychologist, conducted a series of experiments on obedience to authority starting in 1963.
- Experimenter (E) orders the teacher (T), the subject of the experiment, to give severe electric shocks to a learner (L) each time the learner answers a question incorrectly.
- The learner is actually an actor, and the electric shocks are not real, but a prerecorded sound is played each time the teacher administers an electric shock.



http://en.wikipedia.org/wiki/File:Milgram_Experiment_v2.png

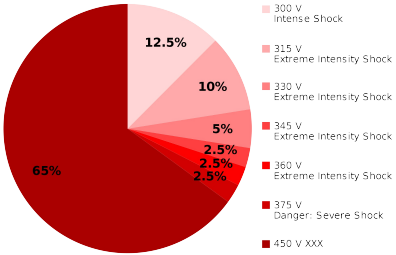

Fake Test Subject
(Urges the participant to keep going)
Experimenter
Subject
(Pretends to feel pain)
(Administers shocks)


PHASE
VOLTAGE ENERGIZER
ATTENUATOR
PULSE FREQUENCY
VOLTAGE
X10
VERY STRONG SHOCK
INTENSE SHOCK
195 VOLTS
255 VOLTS
315 VOLTS
375 VOLTS
435 VOLTS
450 VOLTS


Percentage of participants who delivered voltage
100
90
80
70
60
50
40
30
20
10
0
Slight to moderate shock: 15–135 volts
Strong to very strong shock: 135–255 volts
Intense to extremely intense shock: 255–375 volts
Severe shock: 375-435 volts
XXX: 435-450 volts
Strength of shock

Milgram experiment (cont.)

- These experiments measured the willingness of study participants to obey an authority figure who instructed them to perform acts that conflicted with their personal conscience.
- Milgram found that about 65% of people would obey authority and give such shocks.
- Over the years, additional research suggested this number is approximately consistent across communities and time.



Bernouilli random variables

- Each person in Milgram's experiment can be thought of as a trial.
- A person is labeled a success if she refuses to administer a severe shock, and failure if she administers such shock.
- Since only 35% of people refused to administer a shock, probability of success is $p = 0.35$.
- When an individual trial has only two possible outcomes, it is called a Bernoulli random variable.

Geometric distribution

Dr. Smith wants to repeat Milgram's experiments but she only wants to sample people until she finds someone who will not inflict a severe shock. What is the probability that she stops after the first person?

$$P(1^{st}\ person\ refuses) = 0.35$$

... the third person?

$$P(1^{st}\ and\ 2^{nd}\ shock,\ 3^{rd}\ refuses) = 0.65 \times 0.65 \times 0.35 = 0.65^2 \times 0.35 \approx 0.15$$

... the tenth person?

$$P(9\ shock,\ 10^{th}\ refuses) = \underbrace{0.65 \times \cdots \times 0.65}_{9\ of\ these} \times 0.35 = 0.65^9 \times 0.35 \approx 0.0072$$

Geometric distribution (cont.)

The geometric distribution characterizes the number of trials required to achieve the first success in a sequence of independent and identically distributed (iid) Bernoulli trials, where:

- **Independence:** Outcomes of trials do not affect each other.
- **Identical:** Each trial has the same probability of success.

Geometric Probabilities

Given a success probability p and failure probability $1 - p$, the probability of achieving the first success on the n^{th} trial is given by:

$$P(\text{success on the } n^{th} \text{ trial}) = (1-p)^{n-1}p$$

We denote a random variable $X \sim \text{Geo}(p)$ representing the number of trials until the first success with the probability function:

$$P(X = k) = p(1-p)^{k-1}, \quad k = 1, 2, 3 \ldots$$

Can we calculate the probability of rolling a 6 for the first time on the 6^{th} roll of a die using the geometric distribution? Note that what was a success (rolling a 6) and what was a failure (not rolling a 6) are clearly defined and one or the other must happen for each trial.

(a) no, on the roll of a die there are more than 2 possible outcomes
(b) yes, why not
(c) orangeyes, why not

$$P(6 \text{ on the } 6^{th} \text{ roll}) = \frac{5}{6}^5 \frac{1}{6} \approx 0.067$$

Expected value

How many people is Dr. Smith expected to test before finding the first one that refuses to administer the shock?

The expected value, or the mean, of a geometric distribution is defined as $\frac{1}{p}$.

$$\mu = \frac{1}{p} = \frac{1}{0.35} = 2.86$$

She is expected to test 2.86 people before finding the first one that refuses to administer the shock.

But how can she test a non-whole number of people?

$$EX = \sum_{k=1}^{\infty} k \cdot P(X=k) = \cdots = \frac{1}{p}$$

Expected value and its variability

Mean and standard deviation of geometric distribution

$$\mu = \frac{1}{p} \qquad \sigma = \sqrt{\frac{1-p}{p^2}}$$

- Going back to Dr. Smith's experiment:

$$\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.35}{0.35^2}} = 2.3$$

- Dr. Smith is expected to test 2.86 people before finding the first one that refuses to administer the shock, give or take 2.3 people.
- These values only make sense in the context of repeating the experiment many many times.

We denote a random variable X ⋯ Geo(p) representing the number of trials until the first success with the probability function:

$$P(X=k) = p(1-p)^{k-1}, \quad k=1,2,\cdots$$

$$EX = \frac{1}{p}$$

- Y: # of fails until the first success

$$E(Y) = E(X-1)$$
$$= EX - 1$$

$$P(Y=k) = p(1-p)^k, \quad k=0,1,2,\cdots \qquad Var(Y) = Var(X)$$

$$EY = \frac{1}{p} - 1 = \frac{1-p}{p}$$

$(Y = X-1)$

Binomial distribution

Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will refuse to administer the shock?

Let's call these people Allen (A), Brittany (B), Caroline (C), and Damian (D). Each one of the four scenarios below will satisfy the condition of "exactly 1 of them refuses to administer the shock":

1. Scenario 1: $\frac{0.35}{\text{(A) refuse}} \times \frac{0.65}{\text{(B) shock}} \times \frac{0.65}{\text{(C) shock}} \times \frac{0.65}{\text{(D) shock}} = 0.0961$
2. Scenario 2: $\frac{0.65}{\text{(A) shock}} \times \frac{0.35}{\text{(B) refuse}} \times \frac{0.65}{\text{(C) shock}} \times \frac{0.65}{\text{(D) shock}} = 0.0961$
3. Scenario 3: $\frac{0.65}{\text{(A) shock}} \times \frac{0.65}{\text{(B) shock}} \times \frac{0.35}{\text{(C) refuse}} \times \frac{0.65}{\text{(D) shock}} = 0.0961$
4. Scenario 4: $\frac{0.65}{\text{(A) shock}} \times \frac{0.65}{\text{(B) shock}} \times \frac{0.65}{\text{(C) shock}} \times \frac{0.35}{\text{(D) refuse}} = 0.0961$

The probability of exactly one 1 of 4 people refusing to administer the shock is the sum of all of these probabilities.

$$0.0961 + 0.0961 + 0.0961 + 0.0961 = 4 \times 0.0961 = 0.3844$$

Binomial distribution

The question from the prior slide asked for the probability of given number of successes, k, in a given number of trials, n, ($k = 1$ success in $n = 4$ trials), and we calculated this probability as

$$\#\,of\ scenarios \times P(single\ scenario)$$

- $\#\,of\ scenarios$: there is a less tedious way to figure this out, we'll get to that shortly...
- $P(single\ scenario) = p^k\ (1-p)^{(n-k)}$

  probability of success to the power of number of successes, probability of failure to the power of number of failures

The Binomial distribution describes the probability of having exactly k successes in n independent Bernouilli trials with probability of success p.

Binomial distribution

The Choose Function

The choose function is essential for calculating the combinations of k successes out of n trials:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Binomial Probabilities

Given a success probability p and failure probability $1-p$, for n independent trials and k successes:

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k (1-p)^{n-k}$$

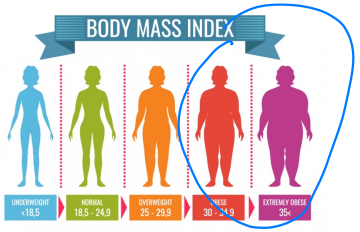
We denote a random variable $X \sim \text{Bino}(n, p)$ with the probability function:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n.$$

A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

- (a) pretty high
- (b) pretty low

Gallup: `http://www.gallup.com/poll/160061/obesity-rate-stable-2012.aspx`, January 23, 2013.



$X \sim Bino(10, 0.262)$

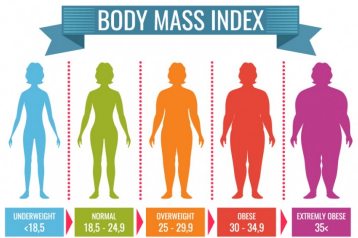
A 2012 Gallup survey suggests that 26.2% of Americans are obese. Among a random sample of 10 Americans, what is the probability that exactly 8 are obese?

(a) $0.262^8 \times 0.738^2$

(b) $\binom{8}{10} \times 0.262^8 \times 0.738^2$

(c) $\binom{10}{8} \times 0.262^8 \times 0.738^2 = 45 \times 0.262^8 \times 0.738^2 = 0.0005$

(d) $\binom{10}{8} \times 0.262^2 \times 0.738^8$



Expected value

A 2012 Gallup survey suggests that 26.2% of Americans are obese.
Among a random sample of 100 Americans, how many would you expect to be obese?

- Easy enough, $100 \times 0.262 = 26.2$.
- Or more formally, $\mu = np = 100 \times 0.262 = 26.2$.

Expected value and its variability

Mean and standard deviation of binomial distribution

$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

- Going back to the obesity rate:

$$\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.262 \times 0.738} \approx 4.4$$

- We would expect 26.2 out of 100 randomly sampled Americans to be obese, with a standard deviation of 4.4.

25 questions

• A, B, C, D E

$X \sim \text{Bino}(25, \frac{1}{5})$

$E(X) = 25 \cdot \frac{1}{5} = 5$

$\sigma = \sqrt{5 \cdot \frac{4}{5}} = 2.$

Unusual observations

Using the principle that observations more than 2 standard deviations from the mean are considered unusual, and utilizing the computed mean and standard deviation, we can determine a range indicating plausible numbers of obese Americans in random samples of 100:

$$26.2 \pm (2 \times 4.4) = (17.4, 35)$$

For observations deemed *very unusual*, the criterion extends to more than 3 standard deviations from the mean:

$$26.2 \pm (3 \times 4.4) = [13.0, 39.4]$$

$\frac{100}{1000} = 10\%$

An August 2012 Gallup poll suggests that 13% of Americans think home schooling provides an excellent education for children. Would a random sample of 1,000 Americans where only 100 share this opinion be considered unusual?

(a) No

(b) Yes

(a) $\frac{10}{100} = 10\%$

(b) $\frac{1000}{10000} = 10\%$

$$\mu = np = 1,000 \times 0.13 = 130$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{1,000 \times 0.13 \times 0.87} \approx 10.6$$

unusual if 130 ± 21.2

very unusual if 130 ± 31.8

Exploring Binomial Distributions

Bin(n, p)

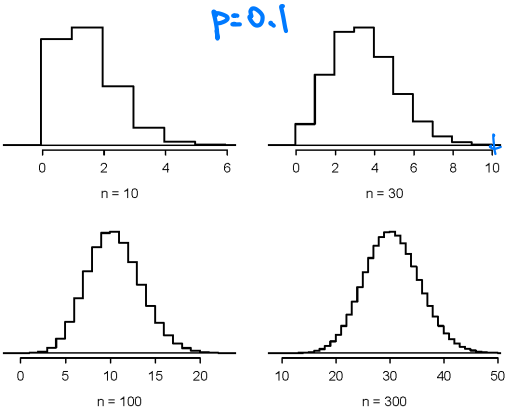
Shapes of Binomial Distributions

Explore using a helpful web applet: navigate to `https://gallery.shinyapps.io/dist_calc/`, and select the "Binomial" from the dropdown menu on the left.

1. Set the number of trials (n) to 20 and the probability of success (p) to 0.15. Observe and describe the shape of the distribution for the number of successes.
2. With p fixed at 0.15, find the minimum sample size (n) that results in a unimodal and symmetric distribution for the number of successes. Each team should submit only one response.
3. Further considerations:
   - Analyze the impact on the distribution's shape when n remains constant and p varies.
   - Examine the changes in the distribution's shape when p remains constant and n varies.

Distributions of number of successes

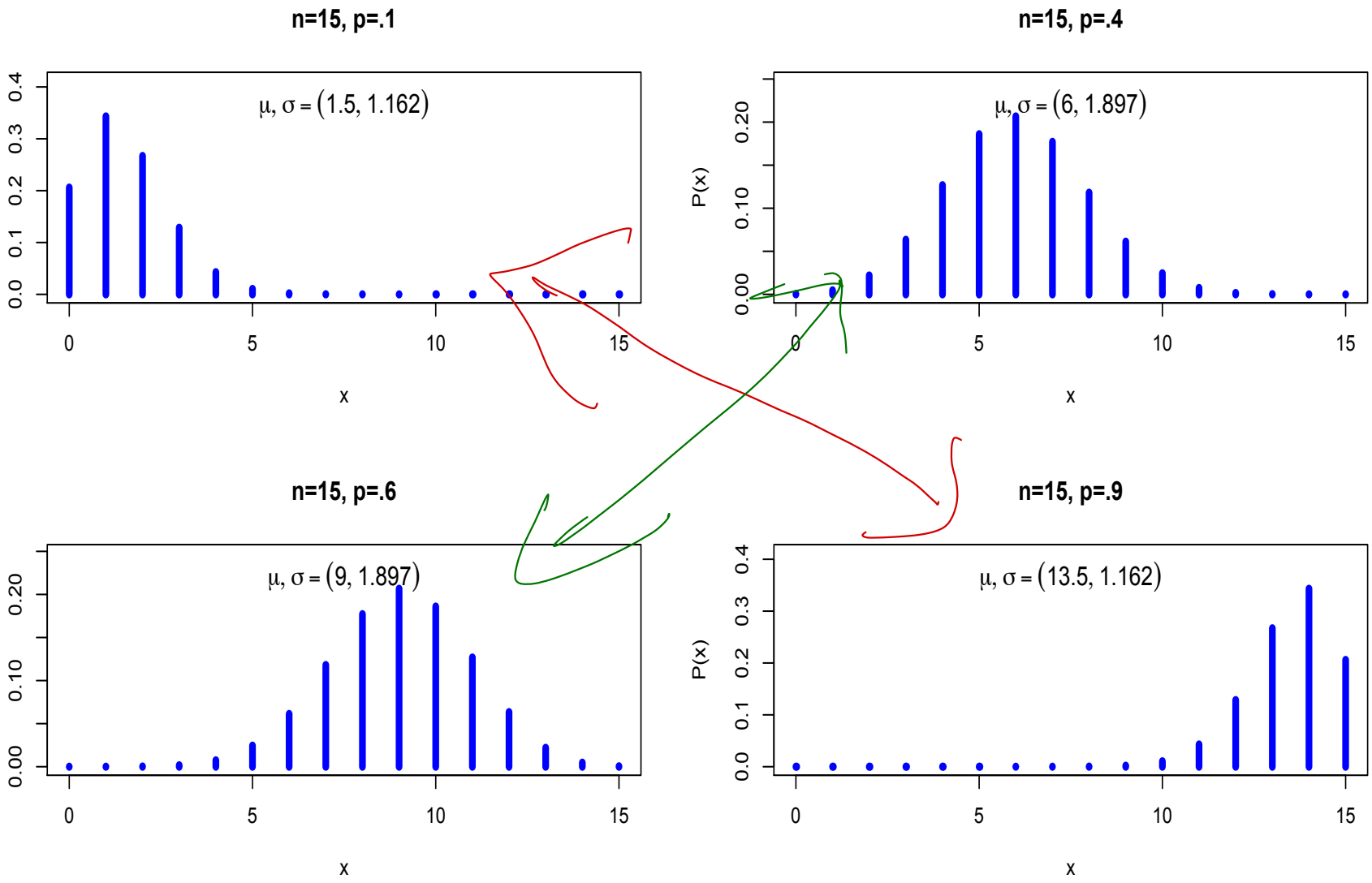
Hollow histograms of samples from the binomial model where $p = 0.10$ and $n = 10$, 30, 100, and 300. What happens as n increases?



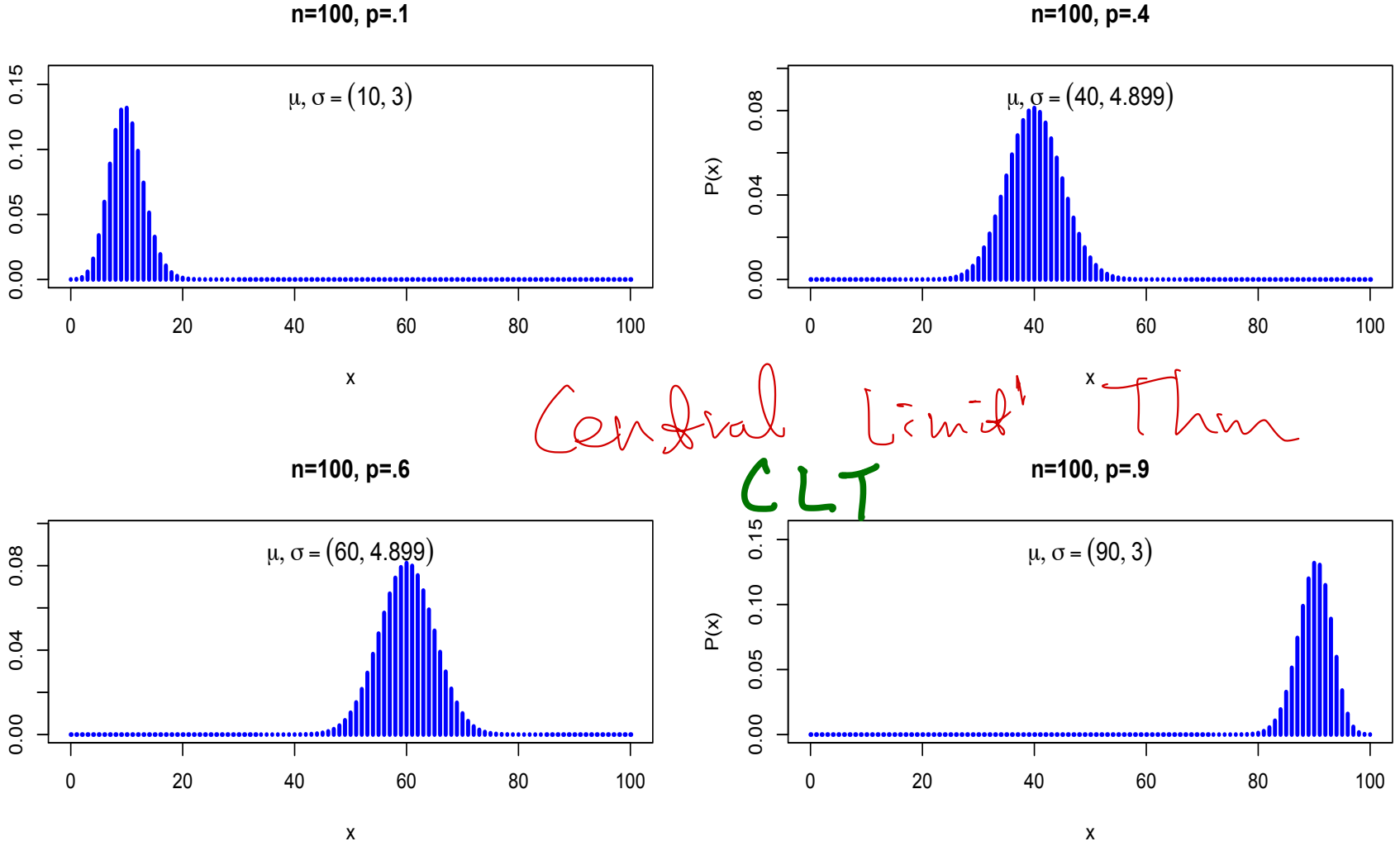
$\binom{30}{11} 0.1^{11} (0.9)^{19}$

≈ 0

二項式機率分佈函數n=15



二項式機率分佈函數n=100



n 大時，機率分佈圖形會趨近於一鐘型曲線

Criteria for Sufficient Sample Size

The sample size is considered sufficiently large for the normal approximation to the binomial distribution if both the expected number of successes and failures are at least 5:

$$np \geq 5 \text{ and } n(1-p) \geq 5$$

Example: For a sample size of $n = 10$ with a success probability of $p = 0.13$:

$$np = 10 \times 0.13 = 1.3$$
$$n(1-p) = 10 \times (1 - 0.13) = 8.7$$

This is NOT a large sample size case.

A study found that approximately 25% of Facebook users are considered power users. The same study found that the average Facebook user has 245 friends. What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users? Note any assumptions you must make.

We are given that $n = 245$, $p = 0.25$, and we are asked for the probability $P(K \geq 70)$. To proceed, we need independence, which we'll assume but could check if we had access to more Facebook data.

$K \sim \text{Bino}(245, 1/4)$

$$\begin{aligned} P(K \geq 70) &= P(K = 70 \text{ or } K = 71 \text{ or } K = 72 \text{ or } \cdots \text{ or } K = 245) \\ &= P(K = 70) + P(K = 71) + P(K = 72) + \cdots + P(K = 245) \end{aligned}$$

This seems like an awful lot of work...

$$\binom{245}{70} p^{70} (1-p)^{175}$$

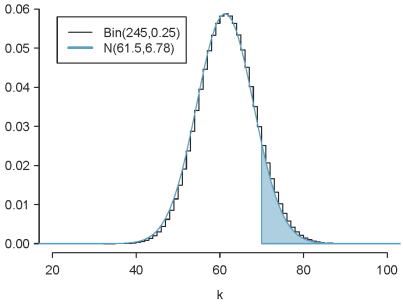
Normal approximation to the binomial

When the sample size is large enough, the binomial distribution with parameters n and p can be approximated by the normal model with parameters $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

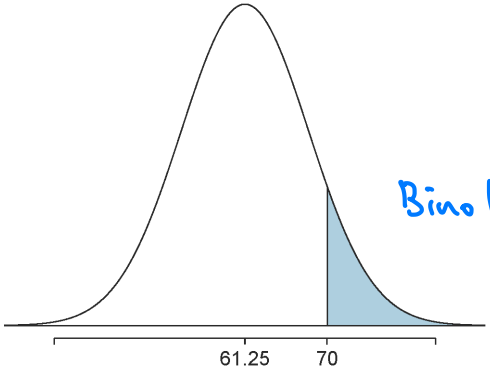
- In the case of the Facebook power users, $n = 245$ and $p = 0.25$.

$$\mu = 245 \times 0.25 = 61.25 \qquad \sigma = \sqrt{245 \times 0.25 \times 0.75} = 6.78$$

- $Bin(n = 245, p = 0.25) \approx N(\mu = 61.25, \sigma = 6.78)$.



What is the probability that the average Facebook user with 245 friends has 70 or more friends who would be considered power users?



$$Z = \frac{obs - mean}{SD} = \frac{70 - 61.25}{6.78} = 1.29$$

$$P(Z > 1.29) = 1 - 0.9015 = 0.0985$$

Bino $(245, 1/4)$

$\mu = 61.25$

$SD = \sqrt{61.25 \cdot 3/4} = 6.78$

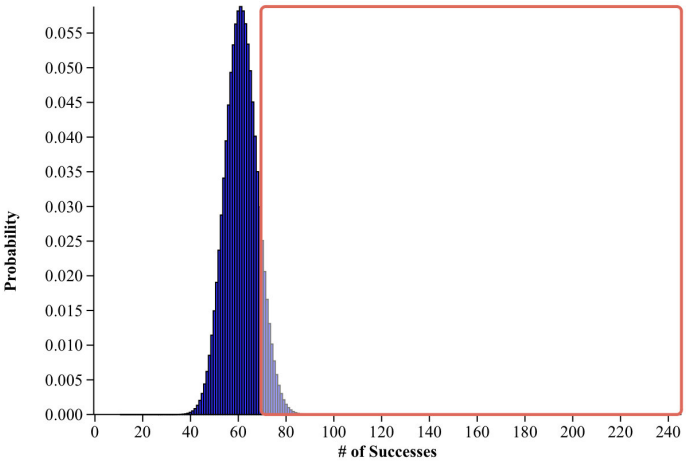
approx.

```
> pnorm(1.29)
[1] 0.9014747
```

exact: pbinom(, ,)

Binomial Distributions

n = 245 p = 0.25 Plot distribution Show normal curve



Mean	Standard Deviation
61.25	6.778

Calculate the probability of at least 70 successes. Go! $P(X \geq 70) = 11.28\%$

-OR-

Calculate the probability of between 12 and 86 successes (inclusive). Go!

Adjust color, rounding, and percent/proportion preferences | Back to menu

Negative binomial distribution

Negative binomial distribution

- The negative binomial distribution describes the probability of observing the r^{th} success on the n^{th} trial.
- The following four conditions are useful for identifying a negative binomial case:
  1. The trials are independent.
  2. Each trial outcome can be classified as a success or failure.
  3. The probability of success (p) is the same for each trial.
  4. The last trial must be a success.

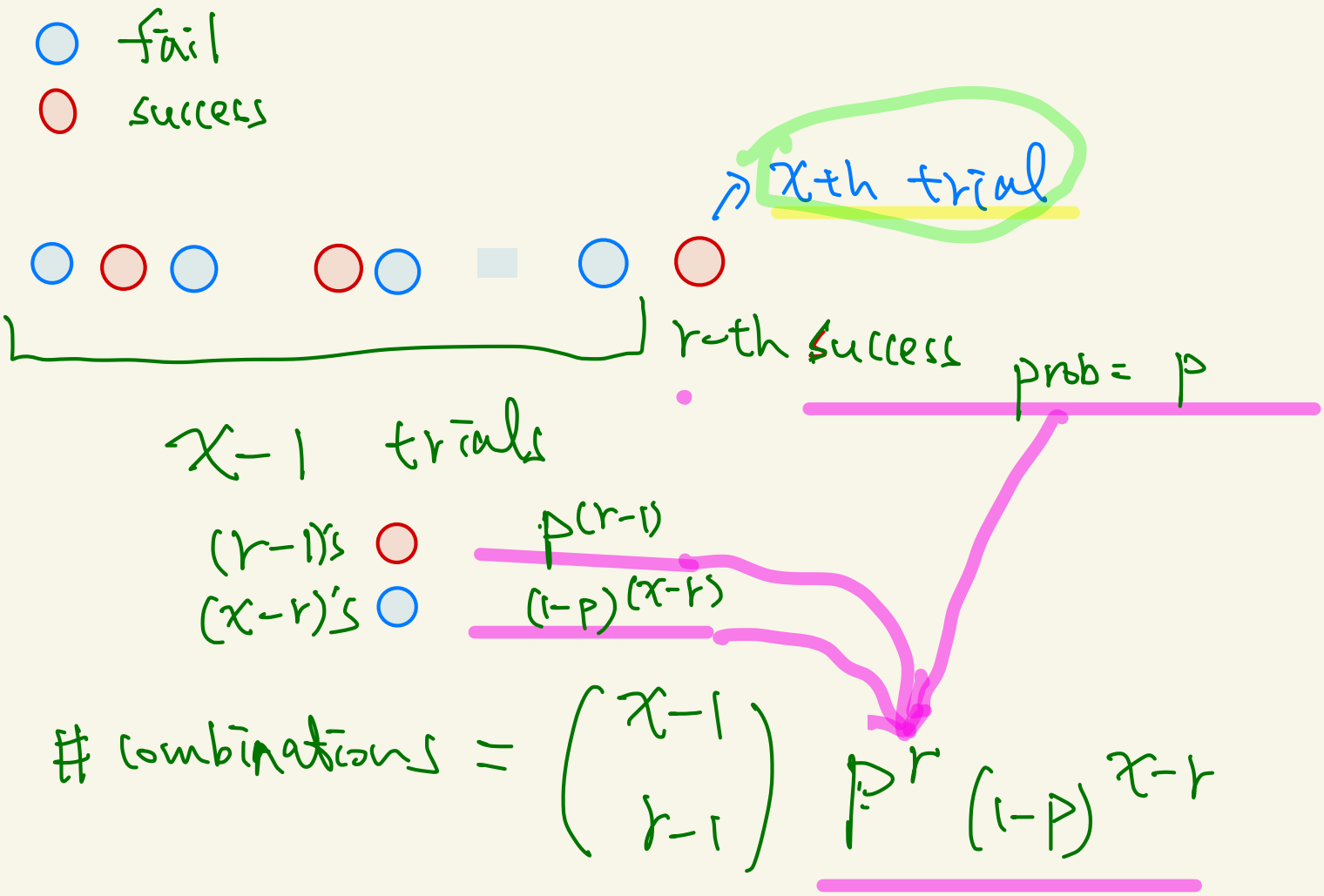
$r=1$ $NB(r, p) \rightarrow Geo(p)$

Negative binomial distribution
We denote a random variable $X \sim NB(r, p)$ with the probability function:

$$P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r},\ x = r, r+1, r+2, \cdots.$$

where p is the probability that an individual trial is a success. The right-hand side represents P(k^{th} success on the x^{th} trial).

$Y = X - r$, # fails to get r-th success


fail
success
xth trial
r-th success
x-1 trials
prob = p
(r-1)'s
(x-r)'s
$p^{(r-1)}$
$(1-p)^{(x-r)}$
combinations = $\binom{x-1}{r-1}$
$p^r (1-p)^{x-r}$

A college student working at a psychology lab is asked to recruit 10 couples to participate in a study. She decides to stand outside the student center and ask every 5^{th} person leaving the building whether they are in a relationship and, if so, whether they would like to participate in the study with their significant other. Suppose the probability of finding such a person is 10%. What is the probability that she will need to ask 30 people before she hits her goal?

Given: $p = 0.10$, $k = 10$, $n = 30$. We are asked to find the probability of 10^{th} success on the 30^{th} trial, therefore we use the negative binomial distribution.

$$\begin{aligned} P(10^{th} \text{ success on the } 30^{th} \text{ trial}) &= \binom{29}{9} \times 0.10^{10} \times 0.90^{20} \\ &= 10,015,005 \times 0.10^{10} \times 0.90^{20} \\ &= 0.00012 \end{aligned}$$

Binomial vs. negative binomial

How is the negative binomial distribution different from the binomial distribution?

- In the binomial case, we typically have a fixed number of trials and instead consider the number of successes.
- In the negative binomial case, we examine how many trials it takes to observe a fixed number of successes and require that the last observation be a success.

Bino(n, p)

NB(r, p)

NB(1, p) = Geo(p)

Practice

Which of the following describes a case where we would use the negative binomial distribution to calculate the desired probability?

(a) Probability that a 5 year old boy is taller than 42 inches.
(b) Probability that 3 out of 10 softball throws are successful.
(c) Probability of being dealt a straight flush hand in poker.
(d) Probability of missing 8 shots before the first hit.
(e) Probability of hitting the ball for the $3rd$ time on the 8^{th} try.

Poisson distribution

Poisson Distribution

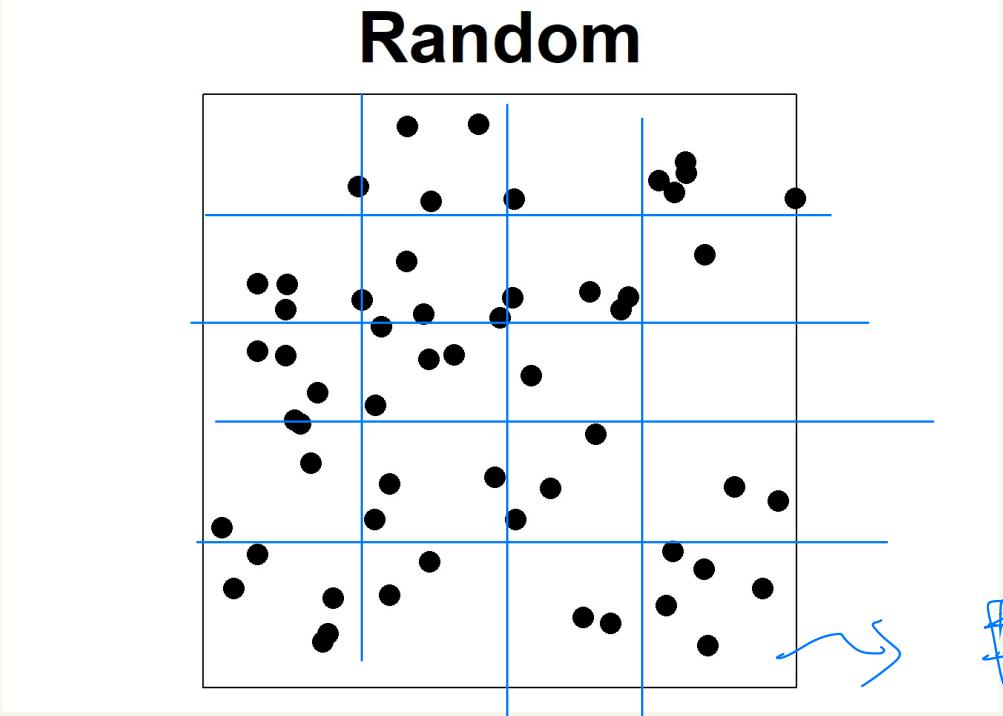
- The Poisson distribution is particularly useful for estimating the frequency of rare events in a large population over a fixed unit of time, assuming independence among individuals.
- The rate (λ) of a Poisson distribution represents the average number of occurrences in a given time period per unit of time.
- This rate allows us to calculate the probability of observing exactly k rare events within a single time unit.

Poisson Distribution Formula
A random variable $X \sim Po(\lambda)$ is described by the probability function:

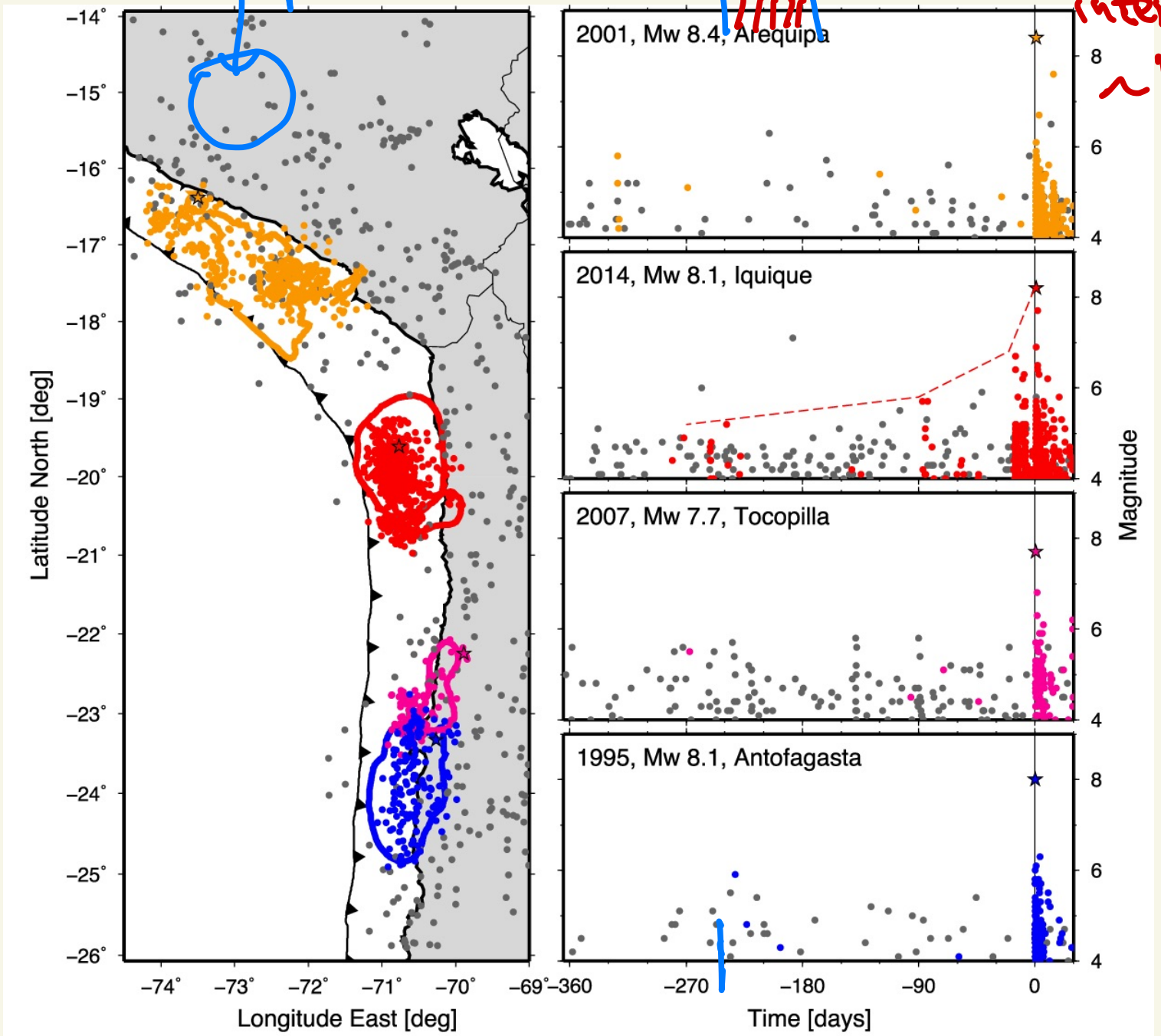
$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, 2, \ldots$$

where e denotes the base of the natural logarithm, the exponential constant approximately equal to 2.718.

Both the mean and variance of X are equal to λ.



# of events
~ Poisson(λa)
a: area of cell.


Poisson
~ # events in an interval
~ Poisson
2001, Mw 8.4, Arequipa
2014, Mw 8.1, Iquique
2007, Mw 7.7, Tocopilla
1995, Mw 8.1, Antofagasta
Latitude North [deg]
Longitude East [deg]
Time [days]
Magnitude
−14°
−15°
−16°
−17°
−18°
−19°
−20°
−21°
−22°
−23°
−24°
−25°
−26°
−74°
−73°
−72°
−71°
−70°
−69°
−360
−270
−180
−90
0
4
6
8

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that in a given week the electricity fails only once.

Given $\lambda = 2$.

$$\begin{aligned} P(\text{only 1 failure in a week}) &= \frac{2^1 \times e^{-2}}{1!} \\ &= \frac{2 \times e^{-2}}{1} \\ &= 0.27 \end{aligned}$$

Suppose that in a rural region of a developing country electricity power failures occur following a Poisson distribution with an average of 2 failures every week. Calculate the probability that on a given day the electricity fails three times.

We are given the weekly failure rate, but to answer this question we need to first calculate the average rate of failure on a given day: $\lambda_{day} = \frac{2}{7} = 0.2857$. Note that we are assuming that the probability of power failure is the same on any day of the week, i.e. we assume independence.

$$\begin{aligned} P(3 \text{ failures on a given day}) &= \frac{0.2857^3 \times e^{-0.2857}}{3!} \\ &= \frac{0.2857^3 \times e^{-0.2857}}{6} \\ &= 0.0029 \end{aligned}$$

Fatal Horse Kicks in the Prussian Army (Bortkiewicz, 1898)

This study is a renowned historical application of the Poisson model, investigating fatal horse kicks in the Prussian army corps between 1875 and 1894. Ten army corps were observed over 20 years, totaling 200 observations. The subsequent table compares the observed data to expected frequencies calculated under the Poisson model, allowing us to evaluate the model's fit.

Number of Deaths	Observed Counts
0	109
1	65
2	22
3	3
4	1
≥ 5	0
total	200

$$\frac{65 + 22\cdot 2 + 3\cdot 3 + 4\cdot 1}{200}$$

$$= \frac{133}{200} = 0.665$$

Comparing Observed and Expected Fatalities

Comparison of observed and Poisson expected numbers of deaths due to horse kicks in the Prussian cavalry.

Number of Deaths	Observed Counts	Poisson Expected
0	109	108.68
1	65	66.28
2	22	20.22
3	3	4.1
4	1	4.64
≥ 5	0	0.1
total	200	200

The close alignment between observed counts and Poisson expectations underscores the model's accuracy in capturing the distribution of rare events.

Approximating Binomial Distribution with Poisson

When n is large and p is small, the Binomial distribution $X \sim Bino(n, p)$ can be approximated by a Poisson distribution with $\lambda = np$.

Example: Predicting Daily AMI Hospitalizations in NYC

- Acute Myocardial Infarction (AMI), commonly known as a heart attack.
- The probability of an individual being hospitalized for AMI on any given day is 0.5 per million.
- New York City has approximately 8 million residents.
- Let Y be the daily number of individuals hospitalized for AMI next Monday. Calculate $P(Y = 0)$, $P(Y = 1)$, and $P(Y = 2)$.
- **Solution:** Since the distribution of Y approximates $\text{Po}(4)$, it yields $P(Y = k) \approx \frac{e^{-4}4^k}{k!}$ for $k = 0, 1, 2$.

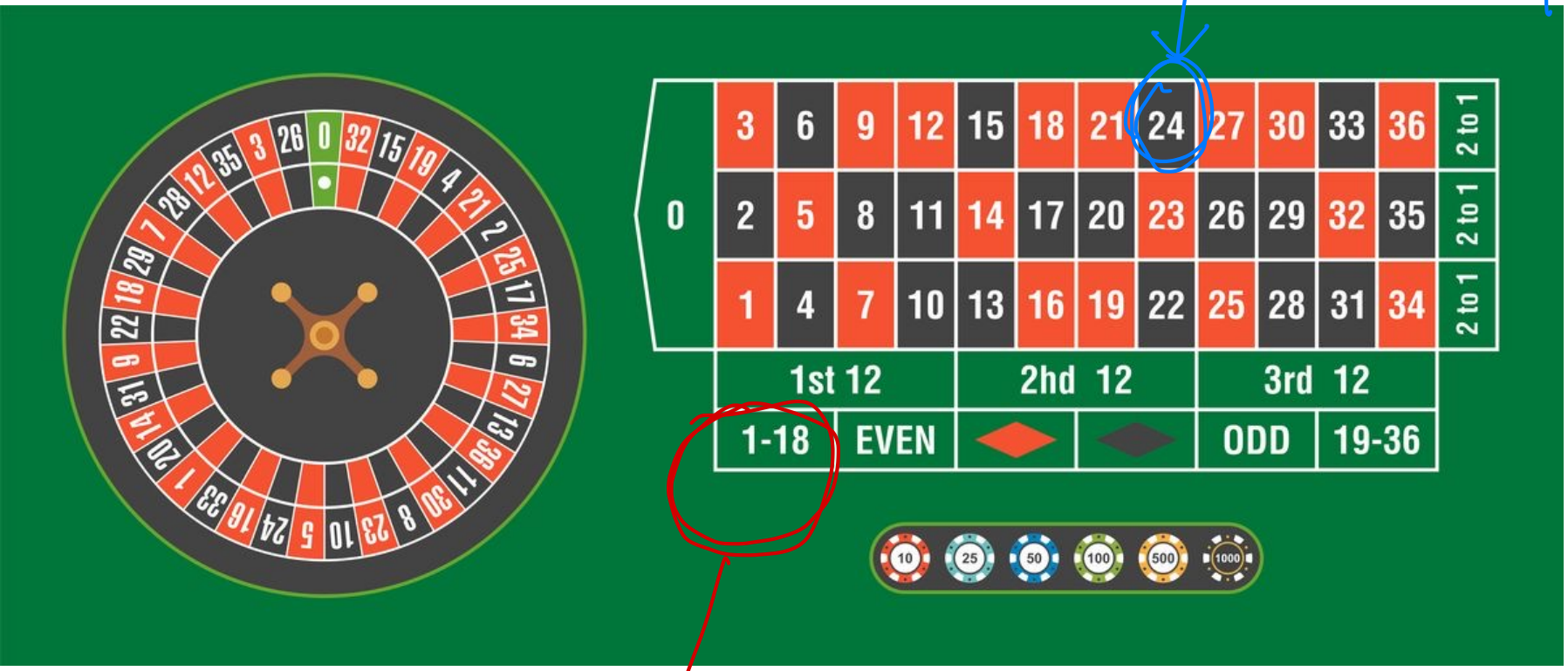
$X \sim B(n, p)$, $np = a$

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \frac{n(n-1)\cdots(n-k+1)}{k!} \left(\frac{a}{n}\right)^k \underbrace{\left(1-\frac{a}{n}\right)^{n-k}}_{\downarrow\ e^{-a}}$$

$$n \to \infty \rightsquigarrow \frac{a^k}{k!} e^{-a}$$

輪盤遊戲



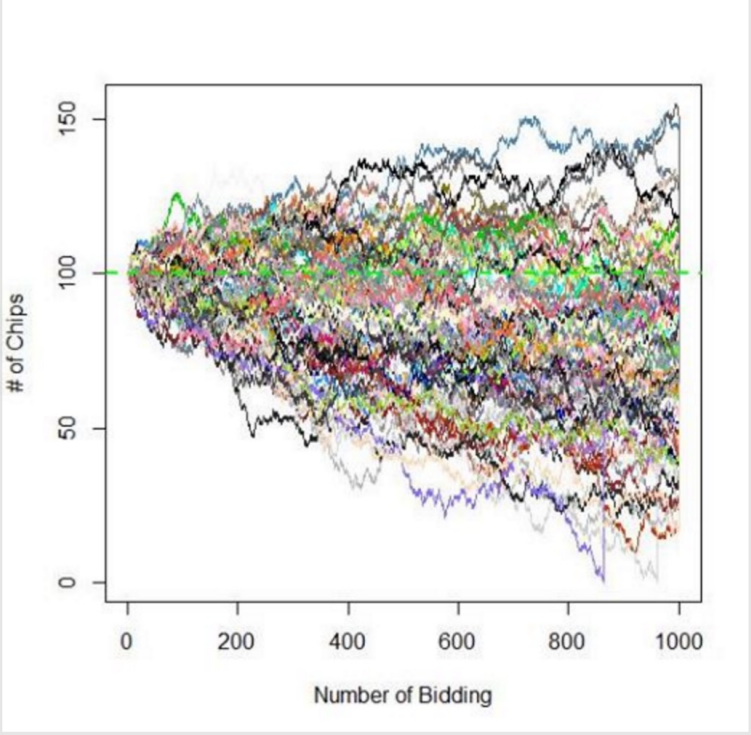
買1賠36

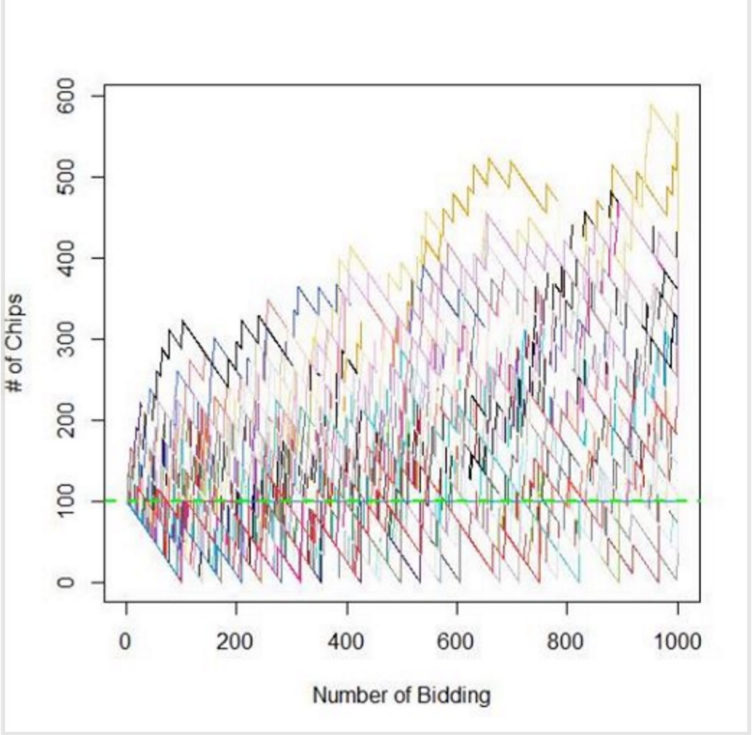
$E(X) = 36 \cdot \frac{1}{37} - 1$

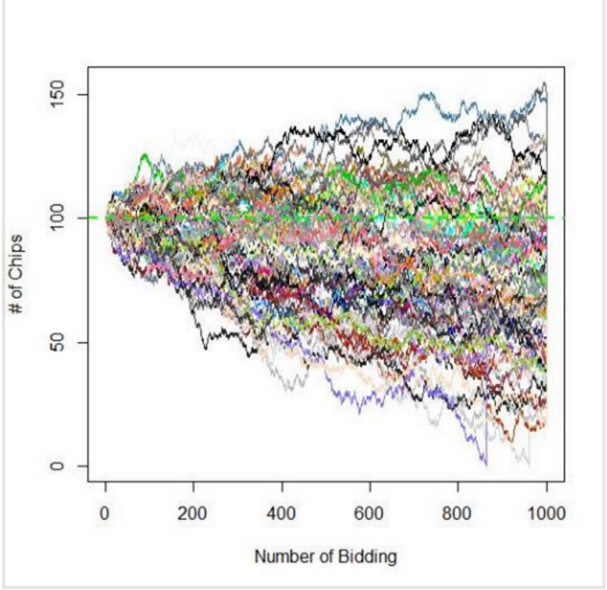
買1賠2，$E(X) = 2 \cdot \frac{18}{37} - 1$

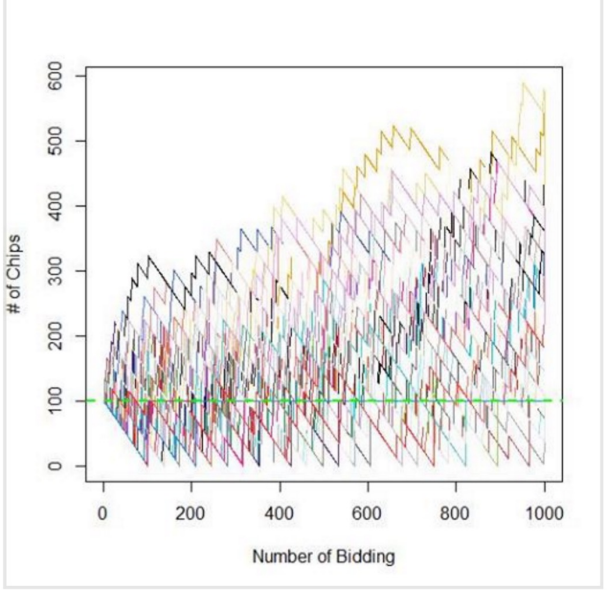
X: reward

- 左圖是壓大小勝率18/37，賠率1 (壓1賠2)；右圖是壓數字勝率1/37，賠率35。兩種壓法期望淨利都是-2.7%。假設有100個籌碼，每次下1個，下面兩張圖為玩1000次，統計100次的累計損益圖，你觀察到什麼？你會採取何種策略？









- 期望值是一樣，右圖破產機會高。
- 右圖之標準差 (風險)遠大於左圖。

每玩一次：

左圖:X~B(1,a), a=18/37, 賠 2X

右圖:Y~B(1,b), b=1/37, 賠 36Y

- $Var(2X) = 2^2a(1-a) \approx 1$ $Var(36Y) = 36^2b(1-b) \approx 34.1$
- 標準差為 1: 5.8

魷魚遊戲

18


02

魷魚遊戲

- 第五關卡是玻璃橋，進行到玻璃橋遊戲時剩16名玩家站在高台上，面前有一座必須走 18 步才能通過的玻璃橋。
- 玩家依序進行，在每一步，玩家都必須在兩塊看似無法區分的玻璃板之間做出選擇，一塊是強化玻璃，另一塊是普通玻璃。強化玻璃可以承受玩家的重量，而普通玻璃只要有人踩上去就會破裂，墜落高台並且掛掉。
- 單純而論，就是每步為二選一猜猜看的運氣考驗。
- 存活率跟出場順序有很大關係。前面犯錯的後面不會再犯。

魷魚遊戲

- 排序一的玩家存活機率是連續 18 步正確，此機率是 $(0.5)^{18}$ ，大約是二十六萬分之一。
- 今16個玩家參與，但只有3個玩家存活的機率為何？

魷魚遊戲

- 16個玩家參與，但只有3個玩家存活的機率為何？
- 只有3個玩家存活，表示有13人陣亡，每次陣亡代表踏錯玻璃（失敗、如同硬幣反面）。
- 由於共有18步玻璃階（投擲硬幣），13人陣亡等同於13次失敗 (5次成功)。
- 令X 代表踩踏18步玻璃階的成功次數，則

X~Binom(18, ½)

$$P(X=5)=\binom{18}{5}\frac{1}{2^{18}}=0.0327$$

魷魚遊戲

- 16個玩家參與，存活人數≤3的機率為何？
- 令X 代表踩踏18步玻璃階的成功次數，X~Binom(18, ½)。
- 3個存活，表示有13次失敗，X=5
- 2個存活，表示有14次失敗，X=4
- 1個存活，表示有15次失敗，X=3
- 0個存活，表示至少有16次失敗，X ≤2

$$P(X \leq 5)=0.048$$

魷魚遊戲的玩家們運氣實在是不太好啊
#編劇太狠心!

魷魚遊戲

- 進階：
- 16個玩家參與，啊Q排第五順位，則啊Q存活的機率為何？
- 令X_i 代表第i人可前進的步數。
- $P(X_i=k)=(0.5)^k$ k=1, 2, 3,… (Geometric distribution)
- 啊Q存活的機率: P(Y>18), $Y=X_1+X_2+\ldots+X_5$

$$P(Y=k)=\binom{k-1}{4}\frac{1}{2^{k-5}}\frac{1}{2^4}\frac{1}{2}=\binom{k-1}{4}\frac{1}{2^k},\quad k=5,6,7,\cdots$$

Negative binomial distribution

P(Y>18)=0.015; E(Y)=10