# 14

## transforming r.v.s

Basic r.v. transportation – Uber example

P( traffic on route A ) = 0.3

If there is traffic on route A, take route B.
- route A is 7min
- route B is 12min

$W \sim \begin{cases} 7\text{min} & \text{w.p. } 0.7 \\ 12\text{min} & \text{w.p. } 0.3 \end{cases}$ not a Bernoulli b/c Supp[X] ≠ {0,1}

$E[W] = \sum_{x\in Supp[X]} x\,p(x) = 7\text{min}\cdot p(7\text{min}) + 12\text{min}\cdot p(12\text{min}) = 7\text{min}\cdot 0.7 + 12\text{min}\cdot 0.3 = 8.5\text{ min}$

- on average in all my trips the time spend in the taxi is ≈8.5 min
- this exact value pertains to an infinite amount of trips

- Uber charges $0.40/min. What is my expected bill per trip ?

to answer this question let's model a new r.v. B (bill)

$B \sim \begin{cases} 7\text{min}\cdot \$0.4/\text{min} = \$2.80 & \text{w.p. } 0.7 \\ 12\text{min}\cdot \$0.4/\text{min} = \$4.80 & \text{w.p. } 0.3 \end{cases}$

B = $0.40/min · W

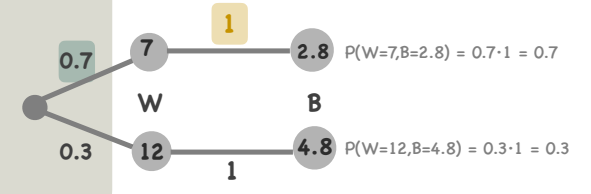
function of W g(W) – depends on W by Def

**transformation of variables** – Math 621

dependent or not ?

$P(W=7\text{min} \mid B=\$2.80) \overset{?}{=} P(W=7\text{min})$

$1 \neq 0.7$



$E[B] = \sum_{x\in Supp[X]} x\,p(x) = \$2.80\cdot p(\$2.80) + \$4.80\cdot p(\$4.80) = \$2.80\cdot 0.7 + \$4.80\cdot 0.3 = \$3.40$

- on average in all my trips (large # of trips) the bill will be ≈ $3.40

BUT… there is another way to look at this.

$E[B] = \$0.40/\text{min}\cdot 8.5\text{min} = \$3.40$

$E[B] = \$0.40/\text{min}\cdot E[W]$

Def: $g(x) = a\,f(x)$ st. $a\in\mathcal{R}$ – **scaling function**

note that B = g(W) = 0.4 W

Is this a coincidence ? Can we generalize this to be valid in all cases ?

$E[Y] \overset{?}{=} a\cdot E[X]$ – a very simple 'scaling function' that takes a function E[X] and scales it by a

$Y = a\cdot X$ | a∈ℛ, X is r.v.

$E[aX] = \sum_{x\in Supp[X]} ax\,p(x)$ | pull the constant a

by 'fake proof' $E[g(x)] = \sum_{x\in Supp[X]} g(x)\,p(x)$

$= a \sum_{x\in Supp[X]} x\,p(x)$ → E[X]

**E[ aX ] = a E[X]**

you can pull the constant through

what if Uber charged a minimum base fee of $3.00 ? Well.. How about another r.v. T

$T \sim \begin{cases} 7\text{min}\cdot \$0.4/\text{min} + \$3 = \$5.80 & \text{w.p. } 0.7 \\ 12\text{min}\cdot \$0.4/\text{min} + \$3 = \$7.80 & \text{w.p. } 0.3 \end{cases}$

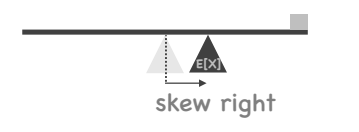
$E[T] = \sum_{x\in Supp[X]} x\,p(x) = \$5.80\cdot p(\$5.80) + \$7.80\cdot p(\$7.80) = \$5.80\cdot 0.7 + \$7.80\cdot 0.3 = \$6.40$

- on average in all my trips (large # of trips) the bill will be ≈ $6.40

BUT… there is another way of looking at this one as well.

$E[T] = \underset{E[B]}{\$3.40} + \$3 = \$6.40$

$E[T] = E[B] + \$3$



Following the example above can we generalize this one as well.

$E[Y] = E[X] + c$ – our T is a very simple function that increases the output of B by 3

$Y = X + c$

$E[X+c] = \sum_{x\in Supp[X]} (x+c)\,p(x)$ | distribute & separate

by 'fake proof' $E[g(x)] = \sum_{x\in Supp[X]} g(x)\,p(x)$

$= \underbrace{\sum_{x\in Supp[X]} x\,p(x)}_{E[X]} + \sum_{x\in Supp[X]} c\,p(x)$ | pull const. c

$= E[X] + c \sum_{x\in Supp[X]} p(x)$ – summing all of the probabilities of the Support of a PMF = 1

remember? $X \sim \begin{cases} 1 & \text{w.p. } 0.6 \\ 0 & \text{w.p. } 0.3 \\ 2 & \text{w.p. } 0.1 \end{cases}$ Σ=1

$= E[X] + c\cdot 1$

**E[ X+c ] = E[X] + c**

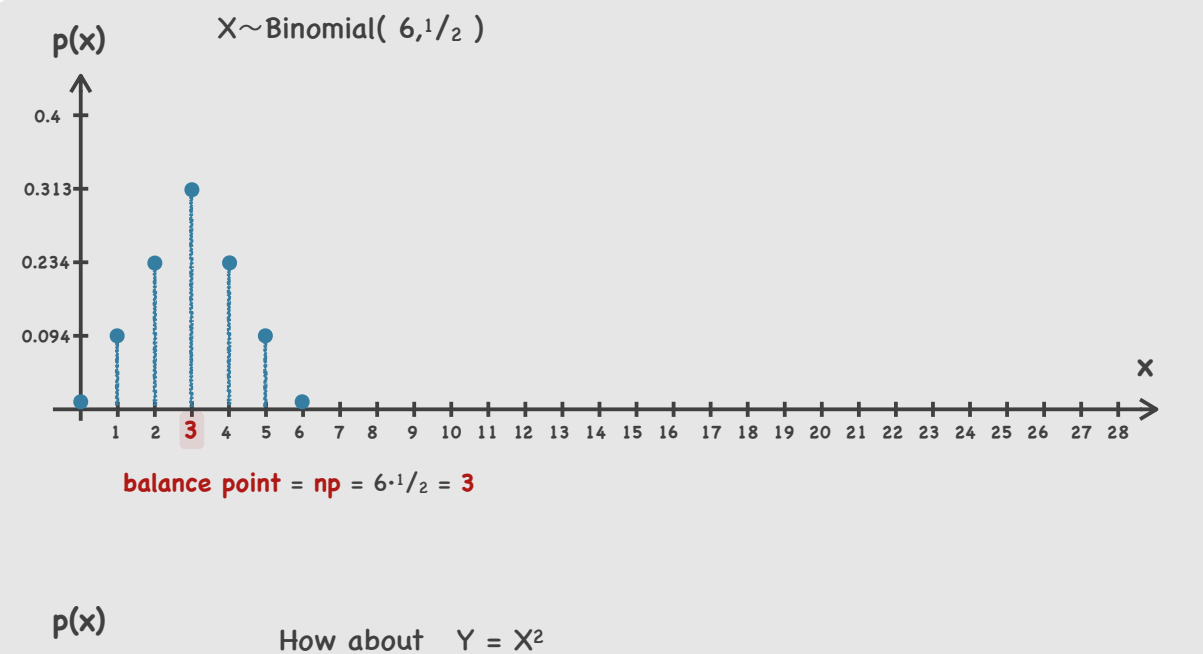
where this E[X] come from ? – E[B] which we generalized as : a E[X]

$= a\,E[X] + c$

Lets rewind the tape…

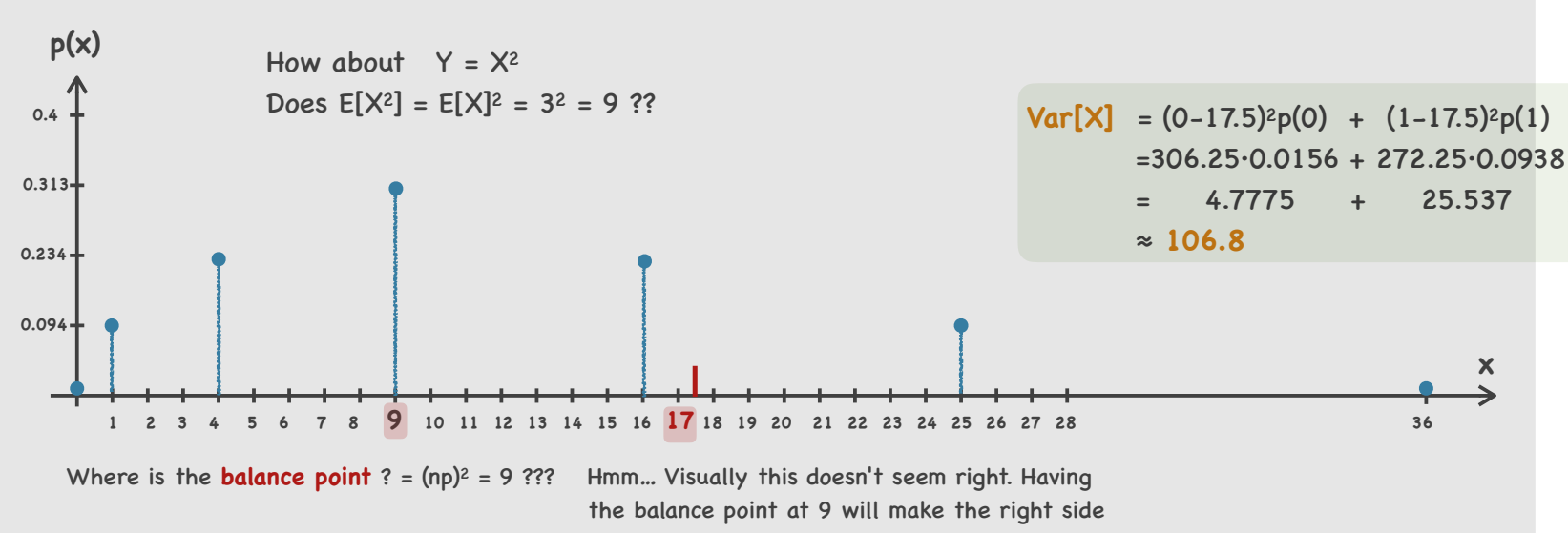
**E[ aX+c ] = a E[X] + c** ← $E[T] \leftarrow E[B] \leftarrow E[W] \leftarrow \overset{r.v.}{W}$ (+c, ·a)

linear function g(x)



**balance point** = np = 6·1/2 = 3

$Var[X] = (0-3)^2p(0) + (1-3)^2p(1) + (2-3)^2p(2) + (3-3)^2p(3) + (4-3)^2p(4) + (5-3)^2p(5) + (6-3)^2p(6)$
$= 9\cdot0.0156 + 4\cdot0.0938 + 1\cdot0.2344 + 0\cdot p(3) + 1\cdot0.2344 + 4\cdot0.0938 + 9\cdot0.0156$
$= .1404 + .3752 + .2344 + 0 + .2344 + .3752 + .1404$
$= 1.5$



How about Y = X²
Does E[X²] = E[X]² = 3² = 9 ??

Where is the **balance point** ? = (np)² = 9 ???

Hmm… Visually this doesn't seem right. Having the balance point at 9 will make the right side of the structure tilt downwards.

$E[X^2] = \sum_{x=0}^{6} x^2 \binom{6}{x} (1/2)^6$ – Supp[X] = {0,1,2,3,4,5,6}

$(1/2)^x(1-1/2)^{n-x} = (1/2)^x(1/2)^{n-x} = a^n a^m = a^{n+m}$

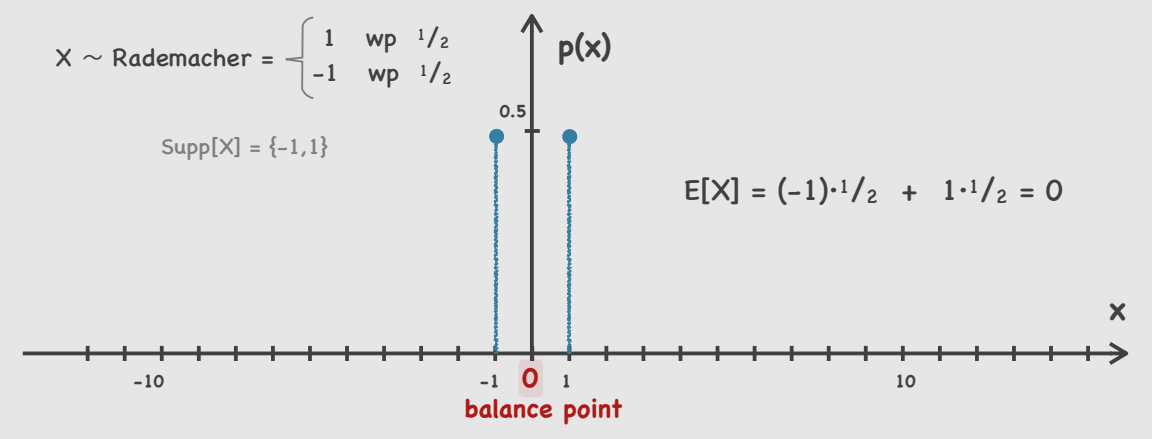
= 17.5 visually it seems a pretty good 'balancing point' but can we be 100% sure that this is correct ?

note that, in general: E[ g(x) ] ≠ g( E[X] )

$Var[X] = (0-17.5)^2p(0) + (1-17.5)^2p(1) + (4-17.5)^2p(2) + (9-17.5)^2p(3) + (16-17.5)^2p(4) + (25-17.5)^2p(5) + (36-17.5)^2p(6)$
$= 306.25\cdot0.0156 + 272.25\cdot0.0938 + 182.25\cdot0.2344 + 72.25\cdot0.3125 + 2.25\cdot0.2344 + 56.25\cdot0.0938 + 342.25\cdot0.0156$
$= 4.7775 + 25.537 + 42.72 + 22.578 + .5274 + 5.276 + 5.3391$
$\approx 106.8$

and now the visually obvious difference between X and X² can be described in a very precise numerical way

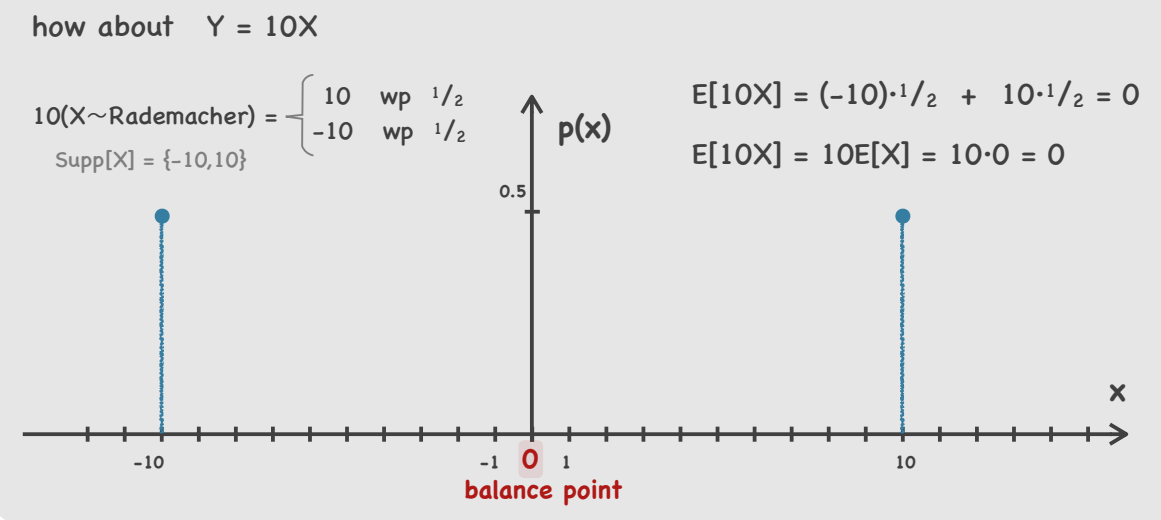
$X \sim \text{Rademacher} = \begin{cases} 1 & \text{wp } 1/2 \\ -1 & \text{wp } 1/2 \end{cases}$ Supp[X] = {-1,1}



$E[X] = (-1)\cdot 1/2 + 1\cdot 1/2 = 0$

how about Y = 10X

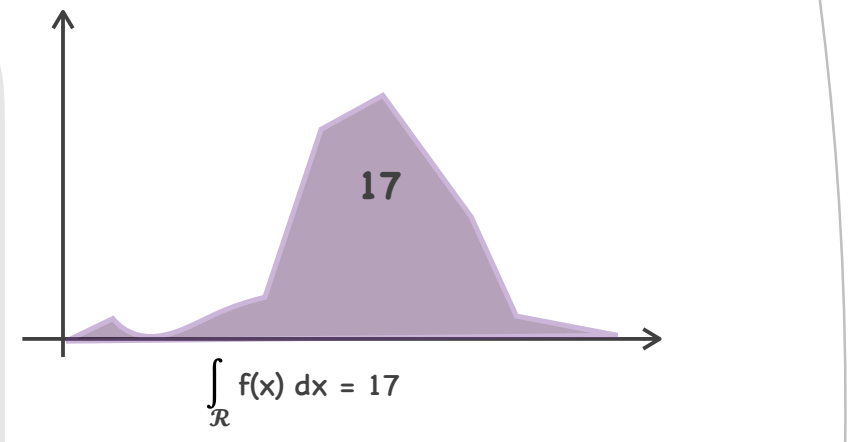
$10(X\sim\text{Rademacher}) = \begin{cases} 10 & \text{wp } 1/2 \\ -10 & \text{wp } 1/2 \end{cases}$ Supp[X] = {-10,10}



$E[10X] = (-10)\cdot 1/2 + 10\cdot 1/2 = 0$

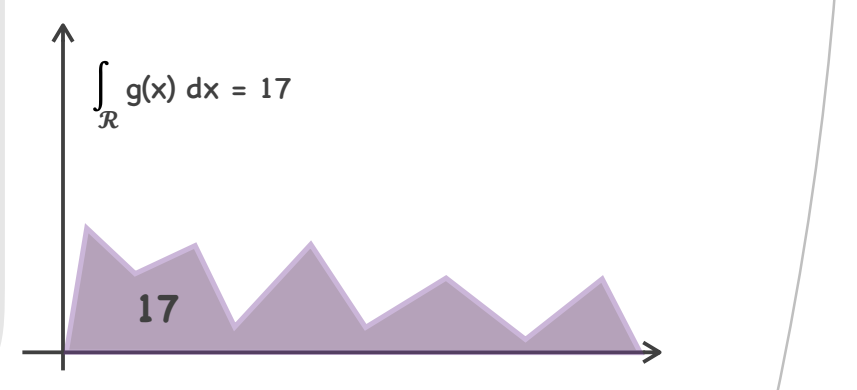
$E[10X] = 10E[X] = 10\cdot 0 = 0$

Two completely different r.v. (distribution of X {1,–1} ≠ distribution of 10X {–10,10}). But if we were to do lots of experiments on both of them, the average ( balancing point ) would be identical. Furthermore, if we were only concerned about that average, then you can also say that those r.v. are the same.



$\int_{\mathcal{R}} f(x)\,dx = 17$

Two totally different functions but if you were to consider only from the perspective of their areas, they are identical.



$\int_{\mathcal{R}} g(x)\,dx = 17$

$Var[X] = (-1-0)^2 p(-1) + (1-0)^2 p(1) = (-1)^2\,1/2 + (1)^2\,1/2 = 1/2 + 1/2 = 1$

$Var[10X] = (-10-0)^2 p(-1) + (10-0)^2 p(1) = (-10)^2\,1/2 + (10)^2\,1/2 = 100\,1/2 + 100\,1/2 = 100$

and now we have a way of to describe numerically the obvious difference between these two r.v.

## Variance

But we can clearly see that they are different. Lets try to express that difference by using another function of a r.v. that captures a metric which describes numerically the difference between them.

Visually speaking we can see that after the transformations were applied both Binomial and Rademacher become more dispersed (less concentrated around the balancing point). Thus our metric should be relative to where the balancing point is and be probability weighted.

theory of loss functions

'error function'
'loss function'

$e(x,\mu) = x - \mu$ : but we don't care if it's + or – (can you have a cost of a –$1) how much is this book? –$1 ?, +$1 ?, $1

$e(x,\mu) = |x - \mu|$ : **absolute loss function – L1**

$e(x,\mu) = (x - \mu)^2$ : **square error loss – L2** Rademacher : 1² = (–1)² = 1 away from μ

abs. val. does the trick but there is another way to convert – into a +

Lets model a new r.v. $L := (X - \mu)^2$

In the 'long run' how far away is x (realization of X) from its 'balance point' squared ?
To answer this question, lets put L through the function that measures the 'long run' behaviorisms.

we also get an expectation but a different from μ

$E[L] = E[(X-\mu)^2] = \sum_{x\in Supp[X]} (x-\mu)^2\,p(x) \longrightarrow \sigma^2 := Var[X] = \sum_{x\in Supp[X]} (x-\mu)^2\,p(x)$

small sigma; switching back to the standard r.v. notation L→X

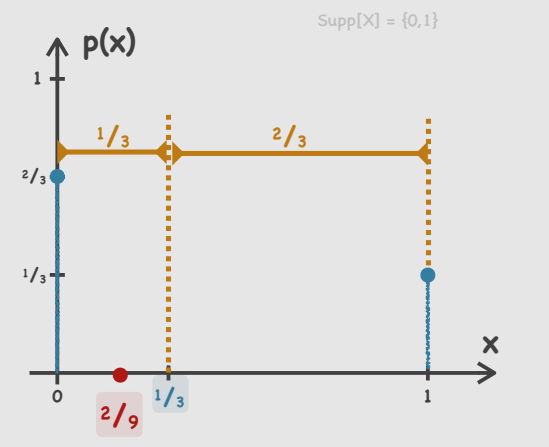
**Variance** aka Deviation

by 'fake proof' using Lebesgue Integral $E[g(x)] = \sum_{x\in Supp[X]} g(x)\,p(x)$

Thus we got another kind of expectation (different from μ) which answers a different question – how far away am I from the balance point μ ? – how dispersed or concentrated is the PMF ?

## Bernoulli Variance

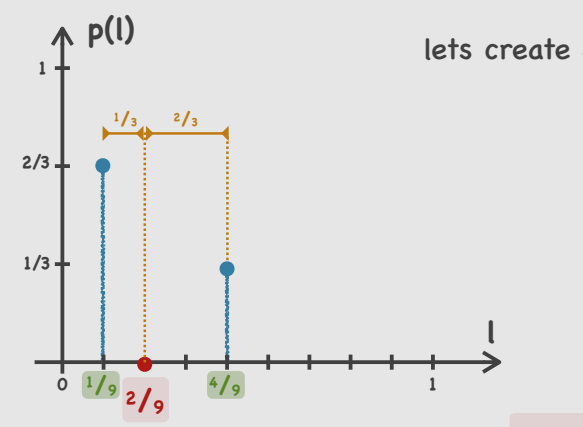
$X \sim \text{Bernoulli}(1/3) = \begin{cases} 1 & \text{wp } 1/3 \\ 0 & \text{wp } 2/3 \end{cases}$ Supp[X] = {0,1}



balance point, E[X] , μ = p = 1/3

σ² – we sum (the distance between each value in the support and μ)² multiplied by the value's corresponding probability

$Var[X] = (0-1/3)^2 (1/3)^0 (2/3)^1 + (1-1/3)^2 (1/3)^1 (2/3)^0 = 1/9\cdot 2/3 + 4/9\cdot 1/3 = 2/27 + 4/27 = 6/27 = 2/9$



lets create a general formula for the Variance of Bernoulli

1/3 · 2/3

$Var[X] = (0-p)^2 (p)^0 (1-p)^1 + (1-p)^2 (p)^1 (1-p)^0$
$= p^2 (1-p) + (1-p)^2 p$
$= (1-p) [ p^2 + (1-p) p ]$
$= (1-p) [ p^2 + p - p^2 ]$
$= (1-p) [ p ]$
$= p (1-p)$

**σ² = Var[X] = p (1-p)**

Previously we analyzed different bets and concluded that they are 'intelligently designed' to have identical 'μ-means' (long-run-average loss per game $.05). Viewing those smilingly disparate bets solely from the perspective of those homogeneous 'μ-means' could lead us to believe that it doesn't really matter which one we should play. However, we just added a new tool to our toolbox which allows us to differentiate between seemingly identical r.v.s . Lets use our newly acquired tool to see if those bets are in fact different.

Roulette in U.S.
total of 38 pockets:
18 black
18 red
2 green 0 , 00

Bet on 'lucky' # 7. Payout is 35:1 (bet $1 win $35)

$X \sim \begin{cases} \$35 & \text{w.p. } 1/38 \\ -\$1 & \text{w.p. } 37/38 \end{cases}$ r.v. model of the payout of this bet

what is the expectation ?

$E[X] = \sum_{x\in Supp[X]} x\,p(x) = \$35\cdot p(\$35) + (-\$1)\cdot p(-\$1) = \$35\cdot 1/38 + (-\$1)\cdot 37/38 = -\$0.053$

Rademacher

Bet on Black. Payout is 1:1 (bet $1 win $1)

$X \sim \begin{cases} \$1 & \text{w.p. } 18/38 \\ -\$1 & \text{w.p. } 20/38 \end{cases}$ r.v. model of the payout of this bet

what is the expectation ?

$E[X] = \sum_{x\in Supp[X]} x\,p(x) = \$1\cdot p(\$1) + (-\$1)\cdot p(-\$1) = \$1\cdot 18/38 + (-\$1)\cdot 20/38 = -\$0.053$

## Standard Error

$Var[X_7] = (\$35 - -\$0.053)^2\,1/38 + (-\$1 - -\$0.053)^2\,37/38$
σ² = 33.207 $² | √ both sides
σ = $5.76
pay attention to the units (extremely important)
but this is no longer 'variance'

law of large numbers
$\bar{X}_7 \to \mu\,(-\$.053)$
$\bar{X}_B \to \mu\,(-\$.053)$

which one gets to μ faster ?
– the one with the lesser distance to it

σ₇ = $5.76
σ_B = $1.00

analogous to: $f(n) = \frac{1}{n}$, $g(n) = \frac{1}{n^{10000}}$; $\lim_{n\to\infty} f(n) = 0$, $\lim_{n\to\infty} g(n) = 0$ – both go to 0 but which one gets there faster?

$Var[X_B] = (\$1 - -\$0.053)^2\,18/38 + (-\$1 - -\$0.053)^2\,20/38$
σ² = .997 $² ≈ 1 $² | √
σ = $1

'Variance – σ²' is an expectation which means it has an interpretation – it's a long run average of something

On the other hand the interpretation of the 'standard deviation – σ' is not so clear! It is not an expectation … it is just a practical strategy to report spread. It will be useful later in statistics.

**standard error or standard deviation**

$\sigma := SE[X] = \sqrt{Var[X]} = \sqrt{\sigma^2}$ √variance

σ – standard error
↑ √
Var[X] – σ² – variance
↑ transformation E[L]
E[X] – μ – mean , balance point
↑
r.v. X

Thus playing 'Lucky 7' will stimulate your hopes for lucrative outcomes far longer than playing 'bet on black'.

©Copyright by Martin G Sendrowicz 2017
All Rights Reserved

Mean, Expectation, Long-Run-Average (aka) : is asking the question: if I do something over and over and over again (law of Large Numbers), what is the average thing that is going to happen ? What is the average of all realizations?

Once we have the expectation

Variance is asking: how far away am I from the long-run average in the squared error sense? What is the average of the squared distance away from the mean? If the answer is a very big number then you are dispersed all over the map.

- any r.v.

$Var[X] := E[(X-\mu)^2]$
$= E[X^2 - 2X\mu + \mu^2]$ – expectation of the Sum is the sum of expectations
$= E[X^2] + E[-2X\mu] + E[\mu^2]$ – hard to compute; expectation of a constant is Degenerate and is always μ²
$= E[X^2] + -2\mu E[X] + \mu^2$ (E[X] = μ)
$= E[X^2] + -2\mu\mu + \mu^2$
$= E[X^2] + -2\mu^2 + \mu^2$
$= E[X^2] + -\mu^2$
$-E[X^2] = -Var[X] - \mu^2$

# E[ X² ]

**E[X²] = Var[X] + μ²**

**E[X²] = σ² + μ²**

vocab:

E[X] – 1st moment – μ
E[X²] – 2nd moment
…
E[X^k] – kth moment

E[ |X-μ| ] – 1st centered moment
E[ (X-μ)² ] – 2nd centered moment – Var[X] – σ² – variance
E[ |X-μ|^k ] – 3rd centered moment

$\frac{E[|X-\mu|]}{\sigma}$ – 1st standardized moment
$\frac{E[(X-\mu)^2]}{\sigma^2} = 1$ – 2nd standardized moment
agnostic to whether it is Left or Right: $Skew[X] = \frac{E[(X-\mu)^3]}{\sigma^3}$ – 3rd standardized moment
slope of the curve: $Kurt[X] = \frac{E[(X-\mu)^4]}{\sigma^4}$ – 4th standardized moment – Kurtosis

remember transformation of variables?
Y = aX+c s.t. a∈ℛ , c∈ℛ

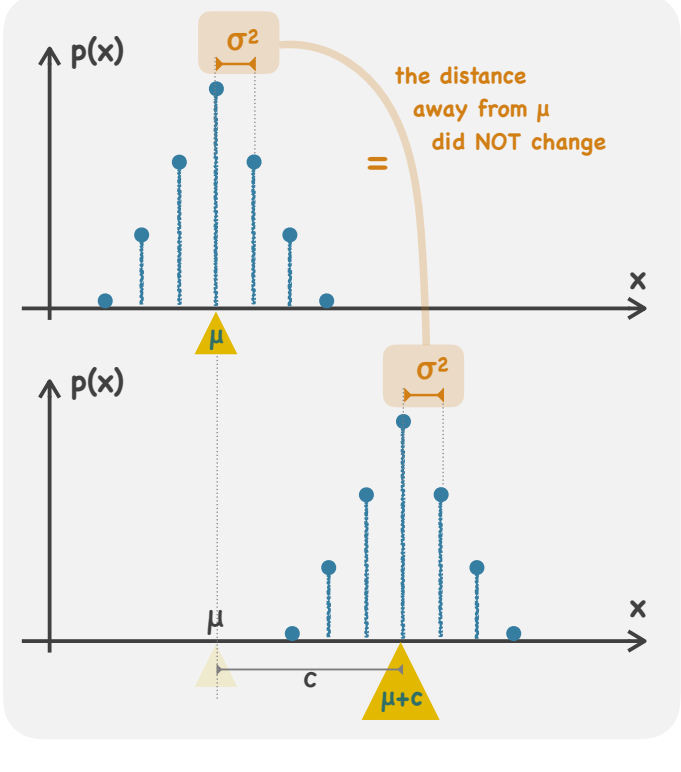
lets do something similar with variance:

Var[X] → Var[X+c] s.t. c∈ℛ

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

the variance did NOT change

**Var[X] = Var[X+c]**



how about:

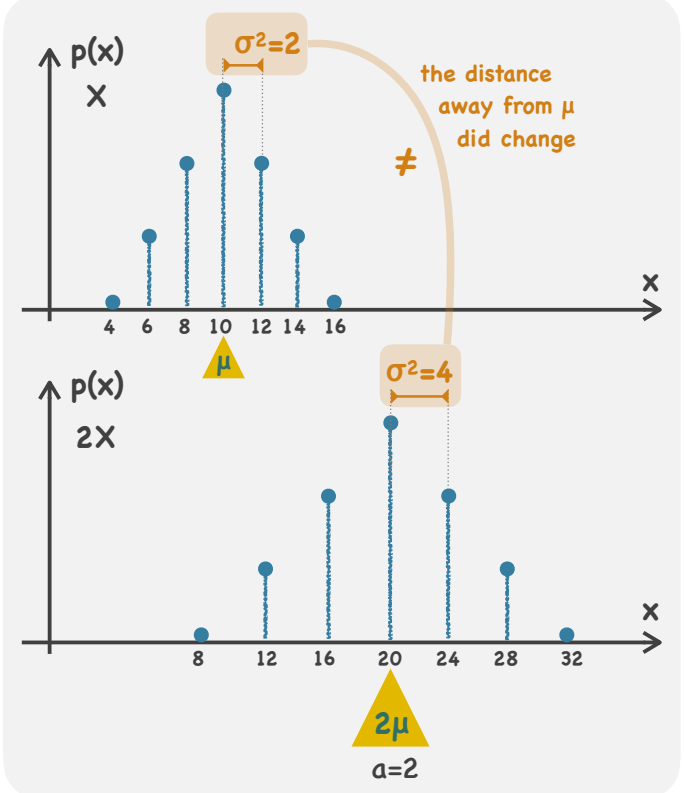
Var[X] → Var[aX] s.t. a∈ℛ

$Var[aX] = E[(aX - a\mu)^2]$
$= E[(a(X-\mu))^2]$
$= E[a^2(X-\mu)^2]$ | pull the const.
$= a^2 \underbrace{E[(X-\mu)^2]}_{\text{aka } Var[X] \text{ aka } \sigma^2 \text{ aka variance}} \to a^2 Var[X]$

**Var[aX] = a² Var[X] = a² σ²**

**Var[aX+c] = a² Var[X] = a² σ²**

**SE[aX+c] = √Var[aX+c] = |a| √Var[X] = |a| √σ² = |a| σ**



SE – standard error – σ:
SE[aX+c] = |a| σ
iid r.v. SE[ X̄ ] = σ/√n

Mean – μ – E[X] – expectation
E[aX] = a E[X]
E[X+c] = E[X] + c
E[aX+c] = a E[X] + c
E[X²] = Var[X]+μ² = σ²+ μ²
E[X₁+…+Xₙ] = E[X₁] +…+ E[Xₙ] – add them up b/c they might be different
E[X₁X₂] = E[X₁] E[X₂] = μ₁ μ₂
identically distributed E[X₁+…+Xₙ] = n E[X] = n μ – don't add just multiply b/c they are all the same
iid r.v. E[ X̄ ] = E[X]

Variance – Var[X] – σ²:
Var[X+c] = Var[X] = σ²
Var[aX+c] = a² σ²
Var[X₁+X₂] = σ₁²+ σ₂²+ E[X₁X₂] – μ₁μ₂ – any r.v.
Var[X₁+…+Xₙ] = Var[X₁]+…+Var[Xₙ] = σ₁²+…+σₙ² – add them up b/c they might be different – independent r.v.
Var[X₁+…+Xₙ] = n Var[X] = n σ² – don't add just multiply b/c they are all the same – iid r.v.
$Var[\bar{X}] = Var[\frac{X_1+X_2+\dots+X_n}{n}] = \frac{\sigma^2}{n}$

## Memorylessness

X ~ Geometric( p ) ← $X_1, X_2, \dots, X_r \overset{iid}{\sim}$ Bernoulli( p )

the 17th experiment was a success given that greater than 10 experiments were a failure $P(X=17 \mid X>10) \overset{?}{=} P(X=7)$ you failed 10 experiments, restarted and started counting again finding a success at the 7th (10+7)

$P(X=7) = (1-p)^6 p$

$P(X=17 \mid X>10) = \frac{P(X=17 \,\&\, X>10)}{P(X>10)} = \frac{P(X=17)}{1-P(10)} = \frac{(1-p)^{16} p}{(1-p)^{10}} = (1-p)^6 p$

**P( X=a+b | X>a ) = P( X=b )**

$P(X=a+b \mid X>a) = \frac{P(X=a+b \,\&\, X>a)}{P(X>a)} = \frac{P(X=a+b)}{1-P(a)} = \frac{(1-p)^{a+b-1} p}{(1-p)^a} = (1-p)^{b-1} p$

'Memorylessness' due to the iid Bernoulli's, if you failed 1000 times, the 1001st begins a new geometric r.v. with parameter p

## Variance cont.

X₁ , X₂ r.v.'s

$Var[X_1 + X_2] = E[((X_1+X_2) - (\mu_1+\mu_2))^2]$ | foil precalc nightmare
…
$= E[X_1^2 + X_2^2 + \mu_1^2 + \mu_2^2 + 2X_1X_2 - 2X_1\mu_1 - 2X_2\mu_2 - 2X_1\mu_2 - 2X_2\mu_1 + 2\mu_1\mu_2]$
$= \underbrace{E[X_1^2]}_{\sigma_1^2+\mu_1^2} + \underbrace{E[X_2^2]}_{\sigma_2^2+\mu_2^2} + \mu_1^2 + \mu_2^2 + 2E[X_1X_2] - 2\mu_1^2 - 2\mu_2^2 - 2\mu_1\mu_2 - 2\mu_1\mu_2 + 2\mu_1\mu_2]$

previously derived E[X²] = σ² + μ²

E[2X₁μ₂] = 2μ₂E[X₁] = 2μ₂μ₁

const E[2μ₁μ₂] = 2μ₁μ₂

E[2X₁μ₁] = 2μ₁E[X₁] = 2μ₁μ₁ = 2μ₁²

| rule: expectation of the Sum is the Sum of the expectations

E[μ²]=μ²

$= \sigma_1^2 + \sigma_2^2 + 2(E[X_1X_2] - \mu_1\mu_2)$

**covariance Cov[X₁,X₂] := E[X₁X₂] – μ₁μ₂**

$= \sigma_1^2 + \sigma_2^2 + 2\,Cov[X_1X_2]$

**Var[ X₁ + X₂ ] = σ₁² + σ₂² + 2 Cov[X₁X₂]**

Covariance is extremely important. I measures how much knowing someones age affects someones vote or height etc… Covariance is the basis of all 'Data Science' as practiced now. It is the basis of all predictions. It is the basis of understanding of relationships between two things. We could practically teach the whole course based on this concept.

lets soften our assumptions about X₁ and X₂

if X₁ and X₂ are **independent** but NOT necessarily identically distributed , what is E[X₁,X₂]

$E[X_1X_2] = \sum_{x_1\in Supp[X_1]} \sum_{x_2\in Supp[X_2]} (x_1x_2)\,p(x_1,x_2)$

$= \sum_{x_1\in Supp[X_1]} \sum_{x_2\in Supp[X_2]} x_1 x_2\, p(x_1)p(x_2)$ – since they're independent lets break them up, pull the constants through the Σ

$= \sum_{x_1\in Supp[X_1]} x_1\,p(x_1) \sum_{x_2\in Supp[X_2]} x_2\,p(x_2)$

$= E[X_1]\;E[X_2]$

**E[ X₁,X₂ ] = E[X₁] E[X₂] = μ₁ μ₂**

notation equivalent to E[ X₁X₂ ]

which means: **Cov[X₁,X₂]** = E[X₁X₂] – μ₁μ₂ = μ₁μ₂ – μ₁μ₂ = **0** (for independent r.v.)

which means: **Var[ X₁ + X₂ ]** = σ₁² + σ₂² + 2 Cov[X₁X₂] = σ₁² + σ₂² + 2·0 = **σ₁² + σ₂²** (for independent r.v.)

in general:
if X₁ , X₂ ,…, Xₙ are **independent** but NOT necessarily identically distributed

$Var[T] = Var[X_1+\dots+X_n] = Var[X_1] +\dots+ Var[X_n] = Var[\sum_{i=1}^n X_i] = \sum_{i=1}^n Var[X_i] = \sum_{i=1}^n \sigma_i^2$

- if independent variance of the sum = the sum of variances – Mnemonic

if X₁ , X₂ ,…, Xₙ are **iid** , what is Var[X₁ + X₂ +…+ Xₙ]

**Var[X₁ + X₂ +…+ Xₙ] = n Var[X] = n σ²** – since they are all identical their Σ collapses to n $\underbrace{\sigma^2+\sigma^2+\dots+\sigma^2}_{n}$

$X_1,\dots,X_n \overset{iid}{\sim} Var[\bar{X}]$

T = X₁ + X₂ +…+ Xₙ

$Var[\bar{X}] = Var[\frac{T}{n}] = \frac{1}{n^2} Var[T] = \frac{1}{n^2}\, n\,\sigma^2 = \frac{\sigma^2}{n}$

# SE[ X̄ ]

$SE[\bar{X}] = \frac{\sigma}{\sqrt{n}}$ ($\frac{\sqrt{\sigma^2}}{\sqrt{n}}$)

one of the most famous formulas in all of statistics

$E[\bar{X}] = E[\frac{X_1+X_2+\dots+X_n}{n}] = E[\frac{X_1}{n}] + E[\frac{X_2}{n}] +\dots+ E[\frac{X_n}{n}]$ | pull the constant through
$= \frac{1}{n}E[X_1] + \frac{1}{n}E[X_2] +\dots+ \frac{1}{n}E[X_n]$ | idd thus identical ·n
$= n\frac{1}{n}E[X]$ | nₛ cancel
$= E[X]$

**E[X̄] = E[X]**

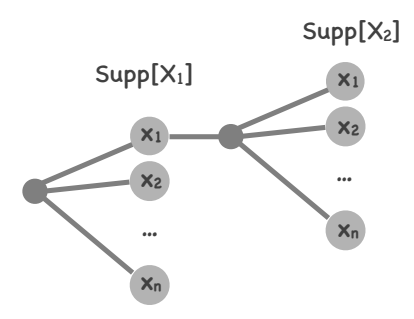
# 15

T = X₁ + X₂

what is the expectation of T? is there a function that represents E[T]?

$E[T] \overset{?}{=} f(E[X_1], E[X_2])$

$= \sum_{t\in Supp[T]} t\,p(t)$ (hard, hard)



note: $\vec{X} = [X_1, X_2]$

T= g(X₁,X₂) = g(X⃗) s.t. g: ℛ²→ ℛ

$E[g(\vec{X})] = \sum_{\vec{x}\in Supp[\vec{X}]} g(\vec{x})\,p(\vec{x}) = \sum_{\langle x_1,x_2\rangle\in Supp[X_1]\cdot Supp[X_2]} g(x_1+x_2)\,p(x_1,x_2)$ – Joint Mass function

$= \sum_{x_1\in Supp[X_1]} \cdot \sum_{x_2\in Supp[X_2]} g(x_1+x_2)\,p(x_1,x_2)$

$E[T] = E[X_1+X_2] = \sum_{\langle x_1,x_2\rangle\in Supp[X_1]\cdot Supp[X_2]} (x_1+x_2)\,p(x_1,x_2) \to \sum_{x_1\in Supp[X_1]} \sum_{x_2\in Supp[X_2]} (x_1+x_2)\,p(x_1,x_2)$ – Joint Mass function

$= \sum_{x_1\in Supp[X_1]} \sum_{x_2\in Supp[X_2]} x_1\,p(x_1,x_2) + \sum_{x_2\in Supp[X_2]} \sum_{x_1\in Supp[X_1]} x_2\,p(x_1,x_2)$ – Σ's switch order, a+b = b+a

$= \sum_{x_1\in Supp[X_1]} x_1 \sum_{x_2\in Supp[X_2]} p(x_1,x_2) + \sum_{x_2\in Supp[X_2]} x_2 \sum_{x_1\in Supp[X_1]} p(x_1,x_2)$

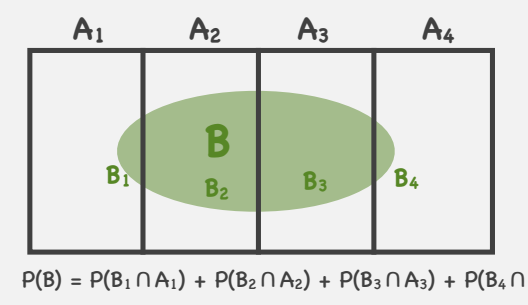
- if X₁ and X₂ are independent r.v.'s , → p(x₁,x₂) = p(x₁) p(x₂) (T = X₁+X₂)

remember? summing all of the probabilities of the Support of a PMF = 1 $X \sim \begin{cases} 1 & \text{w.p. } 0.6 \\ 0 & \text{w.p. } 0.3 \\ 2 & \text{w.p. } 0.1 \end{cases}$ Σ=1

$= \sum_{x_1\in Supp[X_1]} x_1\,p(x_1) \sum_{x_2\in Supp[X_2]} p(x_2) + \sum_{x_2\in Supp[X_2]} x_2\,p(x_2) \sum_{x_1\in Supp[X_1]} p(x_1)$

$= E[X_1]\cdot 1 + E[X_2]\cdot 1$

- consider X₁ , X₂ s.t. Supp[X₁]={ 1,7,19 } , Supp[X₂]={ 5,23,88 }



P(B) = P(B₁∩A₁) + P(B₂∩A₂) + P(B₃∩A₃) + P(B₄∩A₄)

**Law of Total Probability**

Mnemonic 'mutually exclusive' probabilities can be added

|  |  | X₁ 1 | 7 | 19 | Marginal Probability P(X₂) |
|---|---|---|---|---|---|
| X₂ | 5 | 1/15 | 1/3 | 2/15 | 16/30 |
|  | 23 | 1/30 | 1/10 | 1/30 | 5/30 |
|  | 88 | 1/30 | 1/5 | 1/15 | 9/30 |
| P(X₁) Marginal Probability |  | 4/30 | 19/30 | 7/30 | Σ=1 |

P(X₁=1,X₂=5) =

$P(X_1=1) = P(X_1=1,X_2=5) + P(X_1=1,X_2=23) + P(X_1=1,X_2=88) = \sum_{x_2\in Supp[X_2]} p(x_1=1,x_2)$

Σ=1 (rule) $\sum_{x_1\in Supp[X_1]} \sum_{x_2\in Supp[X_2]} p(x_1,x_2) = 1$ – Joint Mass function

similar to: $g(x) = \int_{\mathcal{R}} f(x,y)\,dy$ where did y go? – it got 'integral out'

what do we see here?

- are X₁ and X₂ independent ?

$P(X_1=1) \overset{?}{=} P(X_1=1 \mid X_2=23)$
$4/30 \neq 1/30$

thus NO , they are **dependent** !

- now we need to figure out $\sum_{x_1\in Supp[X_1]} p(x_1,x_2)$ & $\sum_{x_2\in Supp[X_2]} p(x_1,x_2)$

X₂ gets margined out: $p(x_1) = \sum_{x_2\in Supp[X_2]} p(x_1,x_2)$

X₁ gets margined out: $p(x_2) = \sum_{x_1\in Supp[X_1]} p(x_1,x_2)$

this has to hold for all r.v.

$= \sum_{x_1\in Supp[X_1]} x_1 \sum_{x_2\in Supp[X_2]} p(x_1,x_2) + \sum_{x_2\in Supp[X_2]} x_2 \sum_{x_1\in Supp[X_1]} p(x_1,x_2)$

$= \sum_{x_1\in Supp[X_1]} x_1\,p(x_1) + \sum_{x_2\in Supp[X_2]} x_2\,p(x_2)$

$= E[X_1] + E[X_2]$

## E[ T ]
## Mean of r.v.'s Sum

'total r.v.' aka 'sum r.v.'

if T = X₁ + X₂ , **E[T] = E[ X₁+X₂ ] = E[X₁] + E[X₂]**

- general rule for any r.v.'s X₁ , X₂ , … , Xₙ – Mnemonic: expectation of the sum = the sum of expectations

$E[T_n] = E[X_1 + X_2 +\dots+ X_n] = E[\sum_{i=1}^n X_i] = \sum_{i=1}^n E[X_i] = E[X_1] + E[X_2] +\dots+ E[X_n] = \mu_1 + \mu_2 + \dots + \mu_n$

- for any r.v.'s X₁ , X₂ , … , Xₙ – identically distributed (not necessarily independent) meaning all μ's are the same

**E[Tₙ] = E[X₁ + X₂ +…+ Xₙ] = nμ**

## Mean – HyperGeometric

- X ~ **HyperGeometric**( n,K,N )

$E[X] = \sum_{x\in Supp[X]}^{\infty} x\,\underbrace{\frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}}_{PMF}$

can we decompose HyperGeometric into a sum of r.v.'s ?
– yes we can

remember the special case of Hypergeometric ?

X ~ HyperGeometric( 1,1,2 ) (n,K,N) = X ~ Bernoulli( 1/2 )

$p = E[X] = \frac{K}{N}$

so this case can be viewed as:

$n \begin{cases} X_1 \sim \text{Bernoulli}(K/N) \\ X_2 \sim \text{Bernoulli}(K/N) \\ \dots \\ X_n \sim \text{Bernoulli}(K/N) \end{cases}$

T = X₁ + X₂ +…+ Xₙ s.t. X₁ ,…, Xₙ are identically distributed but NOT independent

4 cups with 3 coins in each. There is only 1 marked coin among all those 12 coins. Thus the probability of finding the marked one in a cup depends on it's presence within the other cups (b/c once the marked coin is found no-one else can have it; thus their probability of finding the marked one drops to 0). Each draw viewed independently is a Bernoulli, but viewed as a collective they form a Hypergeometric model. The fact that the draws are dependent is really not important here.

**E[X] = n K/N**

as established earlier…

- X ~ Binomial( n,p ) = $\underbrace{X_1,X_2,\dots,X_n}$ $\overset{iid}{\sim}$ Bernoulli(p)
E[X] = np — $\underbrace{E[X_1]=p,\; E[X_2]=p ,\dots,\; E[X_n]=p}_{n}$

remember? expectation of a Bernoulli r.v. = p

# 16

## Variance – Binomial

X~Binomial( n,p ) = $\underbrace{X_1 , X_2 ,\dots, X_n}_{n} \overset{iid}{\sim}$ Bernoulli( p )

Var[Xₙ] = p (1-p) = σ²

**Var[X] = n p(1-p)**

## Variance – Geometric

$Var[Y] = E[(Y-\mu)^2]$
$= E[Y^2] - \mu^2$ – (previously proved) E[X²] = σ² + μ² = Var[X] + μ²
$= E[Y^2] - 1/p^2$ – (previously proved) geometric μ = 1/p

$E[Y^2] = \sum_{x\in Supp[X]}^{\infty} y^2\,p(x) \to E[Y^2] = \sum_{y=1}^{\infty} y^2 (1-p)^{y-1} p$ – Supp[X]={1,…}

PMF: Y~Geometric( p ) = (1-p)^{y-1} p

(previously proved) $E[Y] = \sum_{y=1}^{\infty} y\,(1-p)^{y-1} p = \frac{1}{p}$

let z=y-1 : y=z+1; y=1…∞; z=0…∞

$= \sum_{z=0}^{\infty} (z+1)^2 (1-p)^z p$
$= p [\sum_{z=0}^{\infty} (z^2+2z+1) (1-p)^z]$
$= p [\sum_{z=0}^{\infty} z^2 (1-p)^z + \sum_{z=0}^{\infty} 2z (1-p)^z + \sum_{z=0}^{\infty} (1-p)^z]$
$= p\sum_{z=0}^{\infty} z^2 (1-p)^z + p\sum_{z=0}^{\infty} 2z (1-p)^z + p\sum_{z=0}^{\infty} (1-p)^z$
$= \sum_{z=0}^{\infty} z^2 (1-p)^z p + \sum_{z=0}^{\infty} 2z (1-p)^z p + \sum_{z=0}^{\infty} (1-p)^z p$
$= (1-p) \underbrace{\sum_{z=1}^{\infty} z^2 (1-p)^{z-1} p}_{E[Y^2]} + 2(1-p) \underbrace{\sum_{z=1}^{\infty} z\,(1-p)^{z-1} p}_{E[Y]=1/p} + p \underbrace{\sum_{z=0}^{\infty} (1-p)^z}_{1/p}$

if z=0 the entire thing = 0 so lets start at 1

$= (1-p)\;E[Y^2] + 2(1-p)\;1/p + p\cdot 1/p$

$1\,E[Y^2] = (1-p)\,E[Y^2] + 2(1-p)\,1/p + 1$
$1\,E[Y^2] - (1-p)E[Y^2] = 2(1-p)\,1/p + 1$ | 1=p/p
$1E[Y^2] - 1E[Y^2] + pE[Y^2] = 2(1-p)/p + p/p$
$pE[Y^2] = 2(1-p)+p/p$ | ÷p
$E[Y^2] = 2(1-p)+p/p^2$
$E[Y^2] = 2-2p+p/p^2$
$E[Y^2] = 2-p/p^2$

$= E[Y^2] - 1/p^2$
$= \frac{2-p}{p^2} - 1/p^2$
$= \frac{2-p-1}{p^2} \to$ **$Var[Y] = \frac{1-p}{p^2}$**