Lecture 7

Random Variables

Definition: A **random variable** is a variable whose value is a numerical outcome of a random phenomenon, so its values are determined by chance. We shall use letters such as X or Y to represent a random variable, and x or y to represent a single outcome of the random variable.

Examples:

(1) Toss a coin four times (an example of a **discrete random variable**).

X = # of heads : 0 1 2 3 4

X is a r.v.

(2) Suppose we are interested in the amount of time a customer spends on a McDonald's drive-thru (an example of a **continuous random variable**).

T = time

$T \in [0, 60 \text{ mins}]$

↳ belongs to

Discrete Random Variables

Definition: A **discrete random variable** X has a finite number of possible values. The **probability distribution** of X lists the values and their probabilities:

Value of X : $x_1,\ x_2,\ \dots\ ,\ x_n$

Probability : $p_1,\ p_2,\ \dots\ ,\ p_n$

Ex. Toss a coin twice

X = # of heads

X :	0	1	2
	x_1	x_2	x_3
P :	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$
	p_1	p_2	p_3

The probabilities p_i must satisfy two requirements:

1. $0 \le p_i \le 1$ for each $i = 1, \dots, n$.
2. $p_1 + p_2 + \cdots + p_n = 1$.

$S = \{HH, HT, TH, TT\}$

We find the probability of any event by adding the probabilities p_i of the particular values x_i that make up the event.

$A = \{$ at least one head $\} = \{1, 2\}$

$P(A) = P(1) + P(2) = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}$

Example: A university posts the grade distribution for its courses online. Students in one section of English 210 in the spring 2006 semester received 31% A's, 40% B's, 20% C's, 4% D's, and 5% F's.

X = grade on 4-point scale

	F	D	C	B	A
X:	0	1	2	3	4
p:	0.05	0.04	0.20	0.40	0.31

$$P(\text{B or higher}) = P(X \geq 3)$$
$$= P(3) + P(4)$$
$$= 0.40 + 0.31 = 0.71$$

Example: We toss a coin four times. What is the probability distribution of the discrete random variable X that counts the number of heads?

Assumptions:

- Coin is fair
- Coin has no memory, i.e. outcomes are independent

$$P(HHHH) = P(H)P(H)P(H)P(H)$$

$$= \left(\frac{1}{2}\right)^4 = \frac{1}{16}$$

$$P(X=4) = \frac{1}{16}$$

		HTTH		
		HTHT		
	HTTT	THTH	HHHT	
	THTT	HHTT	HHTH	
	TTHT	THHT	HTHH	
TTTT	TTTH	TTHH	THHH	HHHH
$X = 0$	$X = 1$	$X = 2$	$X = 3$	$X = 4$

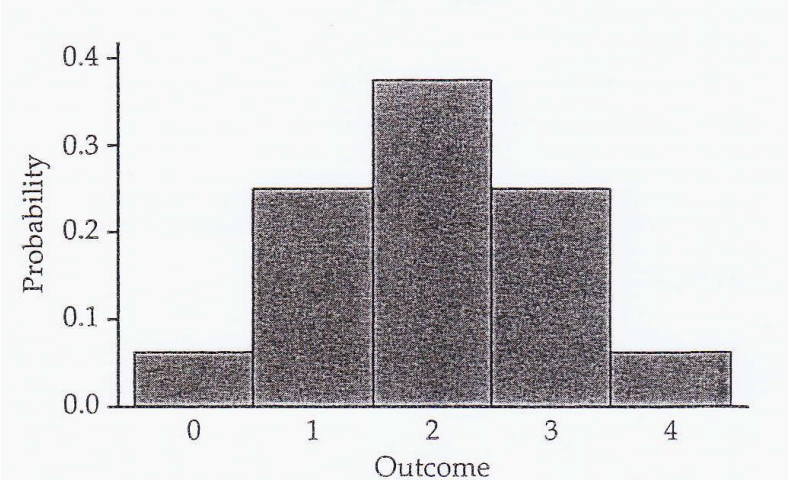
$P(X=3) = 4/16$

$P(X=2) = 6/16$

$P(X=1) = 4/16$

$P(X=0) = 1/16$

X	0	1	2	3	4
P	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$



$P(\text{two or more heads}) = P(X \geq 2)$

$= P(X=2) + P(X=3) + P(X=4)$

$= 6/16 + 4/16 + 1/16 = \frac{11}{16}$

$P(X \geq 1) = 1 - P(X=0) = 15/16$

Mean of a Random Variable

$\bar{x} = \frac{x_1 + \dots + x_n}{n}$ → mean of observations

Example: Most Canadian provinces have government-sponsored lotteries. Here is a simple lottery wager. You choose a three-digit number, 000 to 999. The province chooses a three-digit winning number at random and pays you $500 if your number is chosen.

X = $ your ticket pays you

X:	0	$500
p:	0.999	$0.001 = \frac{1}{1000}$

The long-run average payoff is

$0 \cdot 0.999 + 500 \cdot 0.001 = 0.5$

$\mu_X = E(X) =$ 50 cents ← mean of X

Mean of a discrete random variable:

Suppose X is a discrete random variable whose distribution is

Value of X : $x_1,\ x_2,\ \ldots\ ,\ x_n$

Probability : $p_1,\ p_2,\ \ldots\ ,\ p_n$

$0 \le p_i \le 1$, $\sum_{i=1}^{n} p_i = 1$

Then the **mean** of X is given by

$$\mu_X = x_1p_1 + x_2p_2 + \cdots + x_np_n = \sum_{i=1}^{n} x_ip_i$$

Another notation: $E(X)$ – the **expected value** of X.

Example: If the first digits in a set of data all have the same probability, the probability distribution of the first digit X is then

X	1	2	3	4	5	6	7	8	9
Pr	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9	1/9

$$\mu_X = 1 \cdot \frac{1}{9} + 2 \cdot \frac{1}{9} + \dots + 9 \cdot \frac{1}{9}$$

$$= \frac{1}{9}(1 + 2 + \dots + 9)$$

$$= \frac{1 + 2 + \dots + 9}{9} = 5$$

Ex. Roll a die

X:	1	2	3	4	5	6
P	1/6	1/6	1/6	1/6	1/6	1/6

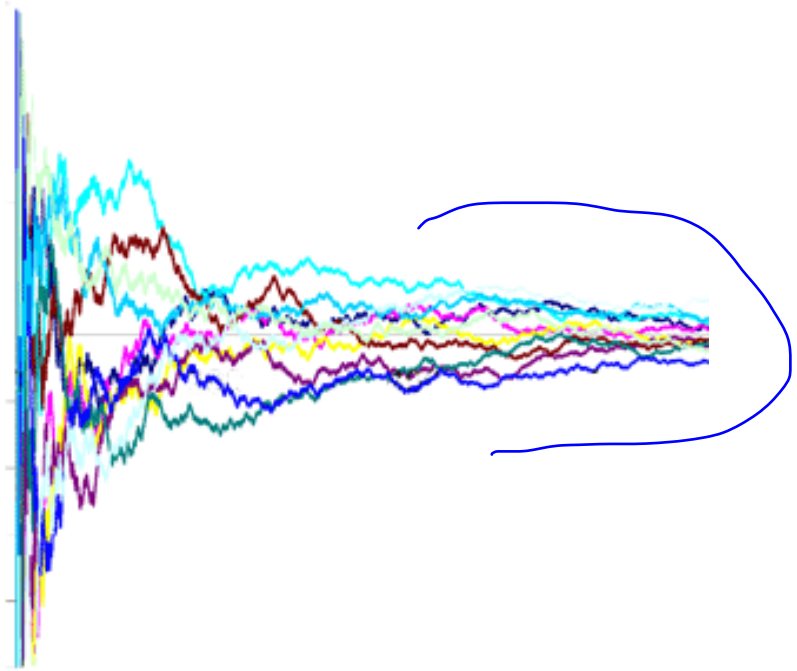
$$\mu_X = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6}$$

$$= \frac{1}{6}(1 + 2 + \dots + 6) = 3.5$$

Statistical Estimation and the Law of Large Numbers (LLN)

μ – parameter, $\bar{X}$ – statistic, $\bar{X}$ is a r.v.

LLN: Draw independent observations at random from any population with finite mean μ. Decide how accurately you would like to estimate μ. As the number of observations drawn increases, the mean $\bar{x}$ of the observed values eventually approaches the mean μ of the population as closely as you specified and then stays that close.



as n increases $\bar{X} \to \mu$

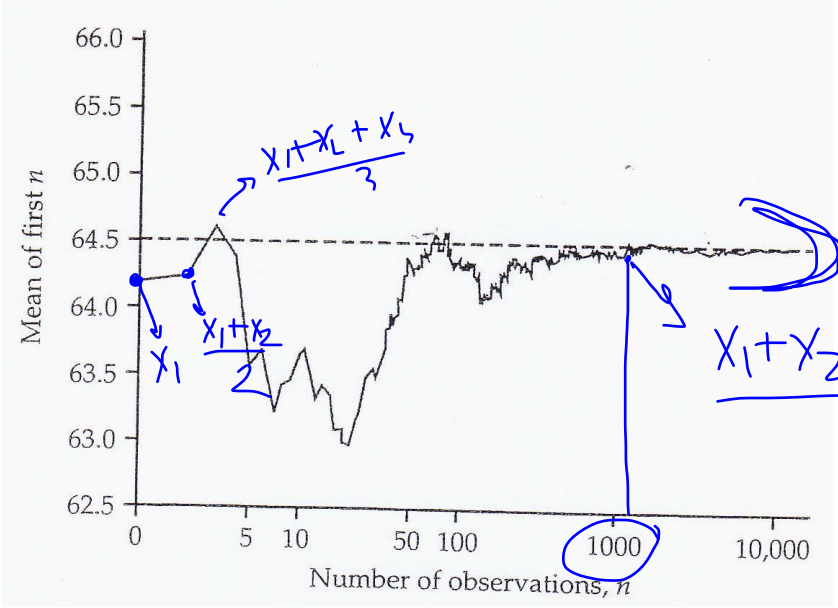
In other words, in the long-run, the average outcome gets close to the distribution mean.

Example: The distribution of the heights of all young women is close to the Normal distribution with mean 64.5 inches and standard deviation 2.5 inches.

$n = 2$, $X_1 = 64.2$, $X_2 = 64.3$

$$\bar{X} = \frac{X_1 + X_2}{2} = \frac{64.2 + 64.3}{2}$$

$$= 64.25$$



$\rightarrow$ close to 64.5 μ

LLN: as n increases

$\bar{X} \rightarrow 64.5 = \mu$

Rules for Means:

1. If X is a random variable and a and b are fixed numbers, then

$$\mu_{a+bX} = a + b\mu_X$$

2. If X and Y are random variables, then

$$\mu_{X+Y} = \mu_X + \mu_Y$$

Why?

Value of X : $x_1,\ x_2,\ \ldots\ ,\ x_n$

Probability : $p_1,\ p_2,\ \ldots\ ,\ p_n$

$a+bX$: $a+bx_1 \ \ldots\ a+bx_n$

P : $p_1 \ \ldots\ p_n$

$$\mu_{a+bX} = \sum (a+bx_i)p_i = \sum (ap_i + bx_i p_i)$$

$$= \sum ap_i + \sum bx_i p_i$$

$$= a\sum p_i + b\sum x_i p_i = a + b\mu_X$$

$\sum p_i = 1$

Example: Linda is a sales associate at a large auto dealership. At her commission rate of 25% of gross profit on each vehicle she sells, Linda expects to earn $350 for each car sold and $400 for each truck or SUV sold. Linda motivates herself by using probability estimates of her sales.

For a sunny Saturday in April, she estimates her car sales as follows:

Cars sold	0	1	2	3
Probability	0.3	0.4	0.2	0.1

Linda's estimate of her truck or SUV sales is

Vehicles sold	0	1	2
Probability	0.4	0.5	0.1

Cars sold	0	1	2	3
Probability	0.3	0.4	0.2	0.1

X = # of cars

$\mu_X = 0 \cdot 0.3 + 1 \cdot 0.4 + 2 \cdot 0.2 + 3 \cdot 0.1 = 1.1$

Y = # of trucks or SUVs

Vehicles sold	0	1	2
Probability	0.4	0.5	0.1

$\mu_Y = 0 \cdot 0.4 + 1 \cdot 0.5 + 2 \cdot 0.1 = 0.7$

Z = Linda's earnings

$Z = \$350 \cdot X + \$400 \cdot Y$

$\mu_Z = \mu_{350X + 400Y} = 350\mu_X + 400\mu_Y$

$= 350 \cdot 1.1 + 400 \cdot 0.7 = \$665$

Variance

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ – sample variance

$Var(X) = \sum (x_i - \mu_X)^2 p_i$

Variance of a discrete random variable: Suppose X is a discrete random variable whose distribution is

Value of X : $x_1,\ x_2,\ \ldots\ ,\ x_n$

Probability : $p_1,\ p_2,\ \ldots\ ,\ p_n$

And μ_X is the mean of X. The **variance** of X is

$$\sigma_X^2 = (x_1 - \mu_X)^2 p_1 + (x_2 - \mu_X)^2 p_2 + \cdots + (x_n - \mu_X)^2 p_n = \sum_{i=1}^{n} (x_i - \mu_X)^2 p_i$$

Another notaion: Var(X).

The **standard deviation** σ_X of X is the square root of the variance.

Example:

Cars sold	0	1	2	3
Probability	0.3	0.4	0.2	0.1

$\mu_X = \sum x_i p_i$

x_i	p_i	$x_i p_i$
0	0.3	0
1	0.4	0.4
2	0.2	0.4
3	0.1	0.3
		$\mu_X = 1.1$

$\sigma_X^2 = \sum (x_i - \mu_X)^2 p_i$

$(x_i - \mu_X)^2 p_i$
$(0-1.1)^2 \cdot 0.3 = 0.363$
$+ (1-1.1)^2 \cdot 0.4 = 0.004$
$+ (2-1.1)^2 \cdot 0.2 = 0.162$
$+ (3-1.1)^2 \cdot 0.1 = 0.361$
$\sigma_X^2 = 0.89$

$\sigma_X = \sqrt{0.89} = 0.94$

Rules for Variances:

1. If X is a random variable and a and b are fixed numbers, then

$$\sigma^2_{a+bX} = b^2\sigma^2_X$$

$$\sigma_{a+bX} = |b|\sigma_X$$

Why?

Value of X : $x_1,\ x_2,\ \ldots\ , x_n$

Probability : $p_1,\ p_2,\ \ldots\ , p_n$

$a+bX$: $a+bx_1$ ... $a+bx_n$

p : p_1 ... p_n

$$\sigma^2_{a+bX} = \sum (a+bx_i - \mu_{a+bx_i})^2 p_i$$

$$= \sum (a+bx_i - (a+b\mu_{x_i}))^2 p_i$$

$$= \sum b^2 (x_i - \mu_{x_i})^2 p_i$$

$$= b^2 \sum (x_i - \mu_{x_i})^2 p_i = b^2\sigma_X^2$$

2. If X and Y are independent random variables, then

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2$$
$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2$$

Note: $\sigma_{X+Y} \neq \sigma_X + \sigma_Y$

$\sigma_{X+Y} = \sqrt{\sigma_{X+Y}^2}$

$= \sqrt{\sigma_X^2 + \sigma_Y^2}$

$\neq \sqrt{\sigma_X^2} + \sqrt{\sigma_Y^2}$

This is the **addition rule for variances of independent random variables**.

3. If X and Y have correlation ρ, then

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$$
$$\sigma_{X-Y}^2 = \sigma_X^2 + \sigma_Y^2 - 2\rho\sigma_X\sigma_Y$$

This is the **general addition rule for variances of random variables**.

Note: The correlation can be found from this formula:

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$$

$$r = \frac{\frac{1}{n-1}\sum (x_i-\bar{x})(y_i-\bar{y})}{s_x s_y}$$

Example: Scores on the Mathematics part of the SAT college entrance exam in a recent year had mean 519 and standard deviation 115. Scores on the Verbal part of the SAT had mean 507 and standard deviation 111. What are the mean and standard deviation of total SAT score?

The correlation between SAT Math and Verbal scores was $\rho = 0.71$.

X = math score, $\mu_X = 519$, $\sigma_X = 115$

Y = verbal score, $\mu_Y = 507$, $\sigma_Y = 111$

Total score $= X + Y$

$$\mu_{X+Y} = \mu_X + \mu_Y = 519 + 507 = 1026$$

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho\sigma_X\sigma_Y$$

$$= 115^2 + 111^2 + 2 \cdot 0.71 \cdot 115 \cdot 111$$

$$= 43,672$$

$$\sigma_{X+Y} = \sqrt{\sigma_{X+Y}^2} = \sqrt{43,672} = 209$$

Sampling Distributions: Counts and Proportions

Example: A sample survey asks 2000 college students whether they think that parents put too much pressure on their children.

We would like to view the responses of these students as representative of a larger population of students who hold similar beliefs. That is, we will view the responses of the sampled students as an SRS from a population.

X = # of 'Yes' → count for outcome of interest

$n = 2000$ → sample size

Assume $X = 840$

$\hat{p} = \frac{X}{n}$ → sample proportion

$= 840/2000 = 0.42$

Binomial Distributions for Sample Counts

The Binomial Setting:

1. There are a fixed number n of observations.
2. The n observations are independent.
3. Each observation falls into one of just two categories (successes and failures)
4. The probability of a success (call it p) is the same for each observation.

Example: Toss a coin 100 times. $n = 100$ $\{H, T\}$

$p = P(H) = \frac{1}{2}$

X = # of heads $X \sim Bin(100, \frac{1}{2})$

Definition: The distribution of the count X of successes in the binomial setting is called the **binomial distribution** with parameters n and p. The possible values of X are the whole numbers from 0 to n.

Notation: $X \sim$ Bin (n, p) or B(n, p)

Example: The probability that a certain machine will produce a defective item is 1/4.

If a random sample of 6 items is taken from the output of this machine, what is the probability that there will be 5 or more defectives in the sample? (The link to Statistical Tables on course website includes table of binomial distribution probabilities. In here, find chance of exactly *k* successes in *n* trials with success probability *p*)

X = # of defectives,

$n = 6$

$p = \frac{1}{4}$

$X \sim Bin(6, \frac{1}{4})$

$P(X \geq 5) = P(X=5) + P(X=6)$

$= 0.0044 + 0.0002$

$= 0.0046$

Example: The financial records of businesses may be audited by state tax authorities to test compliance with tax laws.

It is too time-consuming to examine all sales and purchases made by a company during the period covered by the audit. Suppose the auditor examines an SRS of 150 sales records out of 10,000 available. One issue is whether each sale was correctly classified as subject to state sales tax or not. Suppose that 800 of the 10,000 sales are incorrectly classified. Is the count X of misclassified records in the sample a binomial random variable?

$n = 150$, $p = \frac{800}{10000} = 0.08$

$X \sim Bin(150, 0.08)$

$P(1^{st} \text{ record is bad}) = \frac{800}{10000} = 0.08$

$P(2^{nd} \text{ is bad} \mid 1^{st} \text{ is bad}) = \frac{799}{9999} = 0.079908$

$P(2^{nd} \text{ is bad} \mid 1^{st} \text{ is good}) = \frac{800}{9999} = 0.080008$

Sampling Distribution of a Count:

- Suppose a population contains proportion p of successes.
- If the population is much larger than the sample, the count X of successes in an SRS of size n has approximately the binomial distribution Bin(n, p).
- As a rule of thumb, we will use the binomial sampling distribution of counts when the sample is less than 10% of the population.

Binomial Mean and Standard Deviation

Let $X \sim$ Bin (n, p)

Want: μ_X and σ_X

Let X_i be a random variable that indicates whether the i^{th} observation is success.

X_i : 1 0 → example of

p : p $1-p=q$ Bernoulli r.v.

$\mu_{X_i} = 1 \cdot p + 0(1-p) = p$

$\sigma^2_{X_i} = Var(X_i) = \sum (x_i - \mu_{X_i})^2 p_i = (1-p)^2 \cdot p + (0-p)^2(1-p)$

$= (1-p)p$

Let $X = X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} X_i$

$= \#$ of successes

So $X \sim Bin(n, p)$

$\mu_X = \mu_{X_1 + \ldots + X_n} = \mu_{X_1} + \ldots + \mu_{X_n} =$

$= p + \ldots + p = np$

$\sigma_X^2 = Var(X) = Var(X_1 + X_2 + \ldots + X_n)$

$= Var(X_1) + \ldots + Var(X_n)$

$= p(1-p) + \ldots + p(1-p) = np(1-p) = npq$

$q = 1 - p$

$$\mu_X = np$$

$$\sigma_X = \sqrt{np(1-p)}$$

Example: The Helsinki Heart Study asked whether the anticholesterol drug gemfibrozil reduces heart attacks.

In planning such an experiment, the researchers must be confident that the sample sizes are large enough to enable them to observe enough heart attacks. The Helsinki study planned to give gemfibrozil to about 2000 men aged 40 to 55 and a placebo to another 2000. The probability of a heart attack during the five-year period of the study for men this age is about 0.04. What are the mean and standard deviation of the number of heart attacks that will be observed in one group if the treatment does not change this probability?

X = # of heart attacks

$X \sim Bin(2000, 0.04)$

$\mu_X = np = 2000 \cdot 0.04 = 80$

$\sigma_X = \sqrt{np(1-p)} = \sqrt{2000 \cdot 0.04 \cdot 0.96} = 8.76$

If we observe 50 heart attacks in the treatment group $\Rightarrow$ drug works, because 50 is more than 3 st. dev's below the mean.

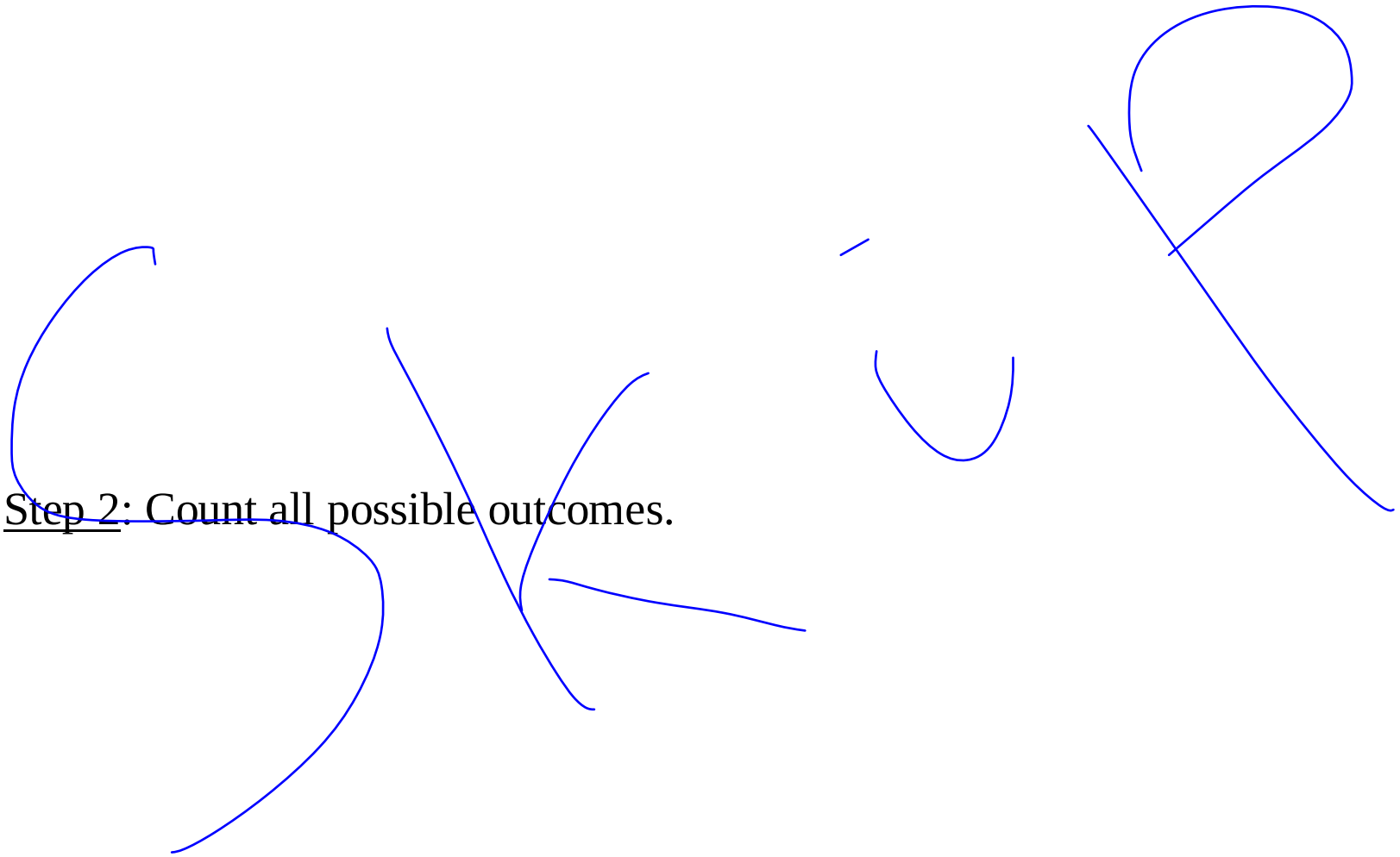
Binomial Formula

Example: Each child born to a particular set of parents has probability 0.25 of having blood type O.

If these parents have 5 children, what is the probability that exactly 2 of them have type O blood?

Let S = success, F = failure

Step 1: Find the probability of a single outcome.



Step 2: Count all possible outcomes.

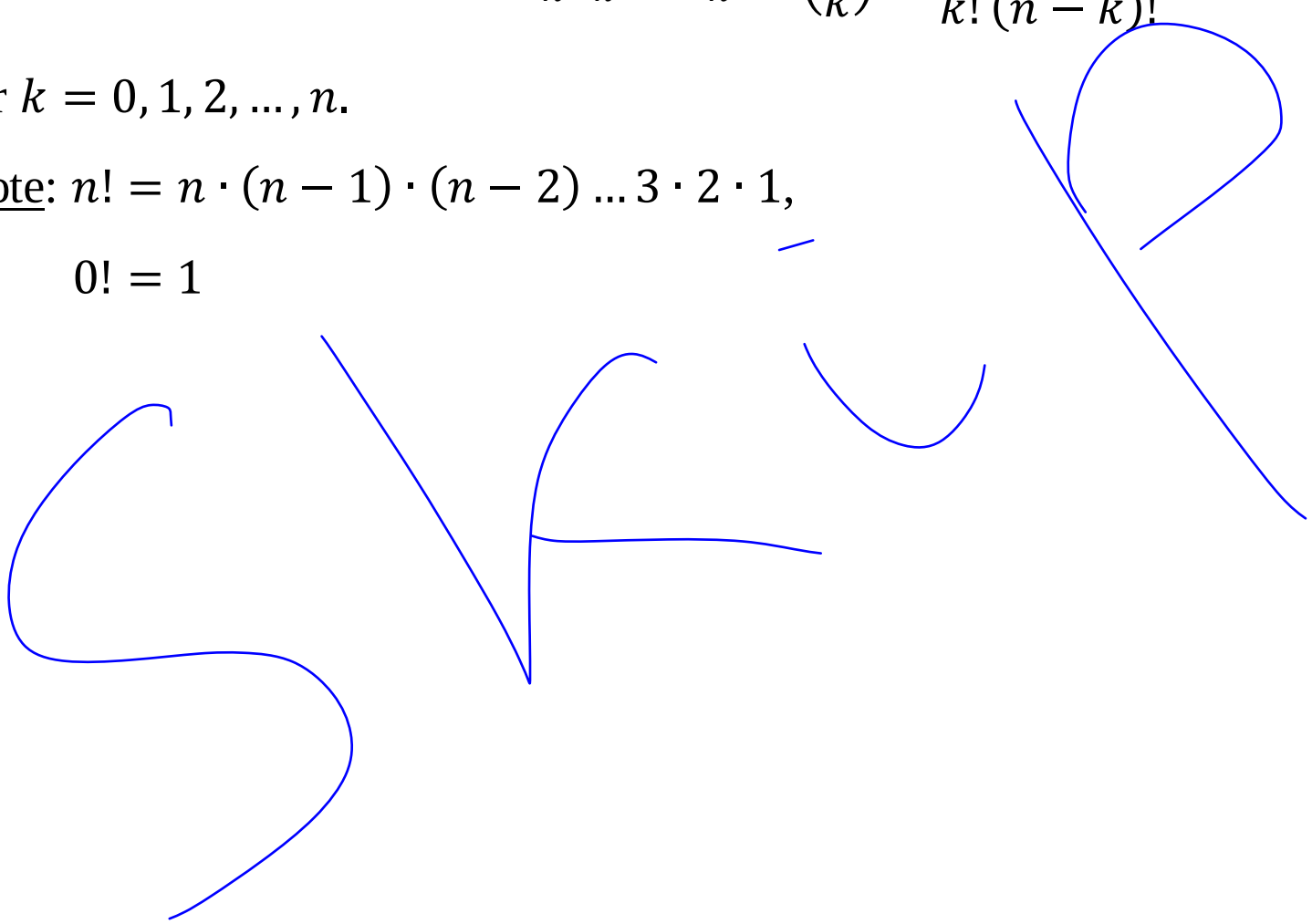
Definition: The number of ways of arranging k successes among n observations is given by the **binomial coefficient**

$$ {}_nC_k = C_n^k = \binom{n}{k} = \frac{n!}{k!\,(n-k)!} $$

for $k = 0, 1, 2, \ldots, n$.

Note: $n! = n \cdot (n-1) \cdot (n-2) \ldots 3 \cdot 2 \cdot 1$,

$0! = 1$



Definition: If X has the binomial distribution Bin(n, p) with n observations and probability p of success on each observation, the possible values of X are 0, 1, 2, ..., n. If k is any of these values, the **binomial probability** is

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

Example: Suppose that the number X of misclassified sales records in the auditor's sample has the Bin(15, 0.08) distribution. What is the probability of at most one misclassified record?

Let's explore the shapes of these binomial distributions (we can use StatCruch for that):

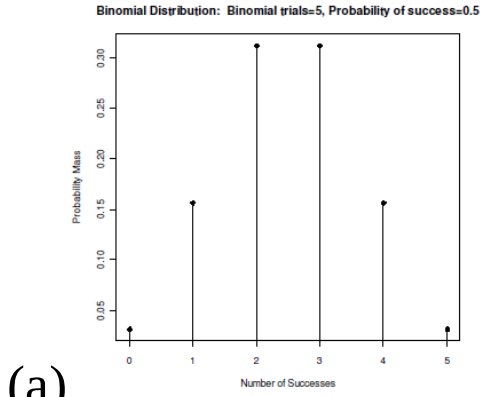
(a) $n = 5$, $p = 0.5$

(b) $n = 20$, $p = 0.9$

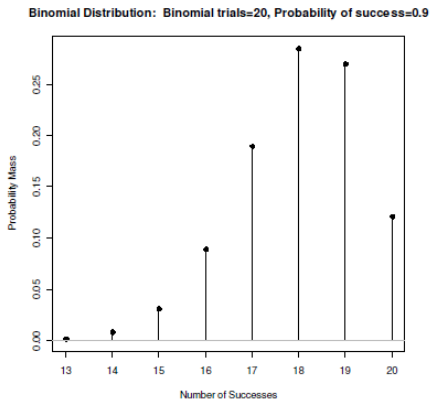
(c) $n = 30$, $p = 0.2$

(d) $n = 500$, $p = 0.4$

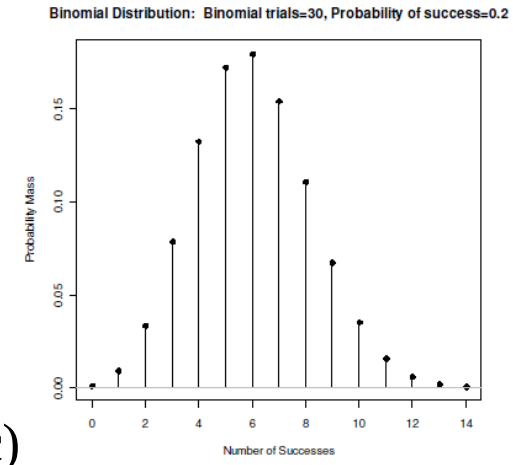
StatCrunch -> Stat -> Calculators -> Binomial



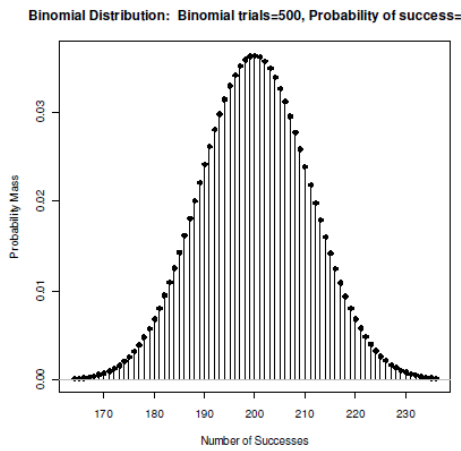
(a)



(b)



(c)



(d)

→ Normal

Which distribution does the last one resemble?

Sample Proportions

X = count, $X \sim Bin(n, p)$

$\hat{p} = \frac{X}{n}$, $0 \le \hat{p} \le 1$

Example: A sample survey asks a nationwide random sample of 2500 adults if they agree or disagree that «I like buying new clothes, but shopping is often frustrating and time-consuming».

X = # of shoppers who would agree

$n = 2500$, $p = 0.6$

Suppose that 60% of all adults would agree if asked this question. What is the probability that the sample proportion who agree is at least 58%?

$$P(\hat{p} \ge 0.58) = P\left(\frac{X}{n} \ge 0.58\right)$$

$$= P(X \ge 0.58 \cdot n) = P(X \ge 0.58 \cdot 2500)$$

$$= P(X \ge 1450) \approx 0.98$$

↓ software

Mean and Standard Deviation of a Sample Proportion

Let $\hat{p}$ be the sample proportion of successes in an SRS of size n drawn from a large population having population proportion p of successes. The mean and standard deviation of $\hat{p}$ are

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{pq}{n}} \qquad q = 1-p$$

Why? $\hat{p} = \frac{X}{n} = \frac{1}{n}X$, $X \sim Bin(n, p)$

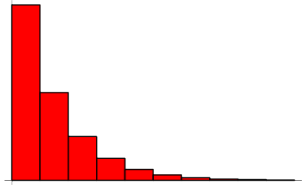
$$\mu_{\hat{p}} = \mu_{\frac{X}{n}} = \frac{1}{n}\mu_X = \frac{1}{n}np = p$$

$$\sigma^2_{\hat{p}} = \sigma^2_{\frac{X}{n}} = \frac{1}{n^2}\sigma^2_X = \frac{1}{n^2}np(1-p) = \frac{p(1-p)}{n}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Note: If $E(\hat{p}) = p$, we say that $\hat{p}$ is an **unbiased estimator** for p.

Geometric Distribution



- Suppose we want to know how long it will take to achieve the first success in a series of Bernoulli trials.
- The model that can tell us this is called the **Geometric probability model**.
- Let p be the probability of success ($q = 1 - p$ is the probability of failure).
- Let X be the number of trials until the first success occurs.
- Then the distribution of X is given by

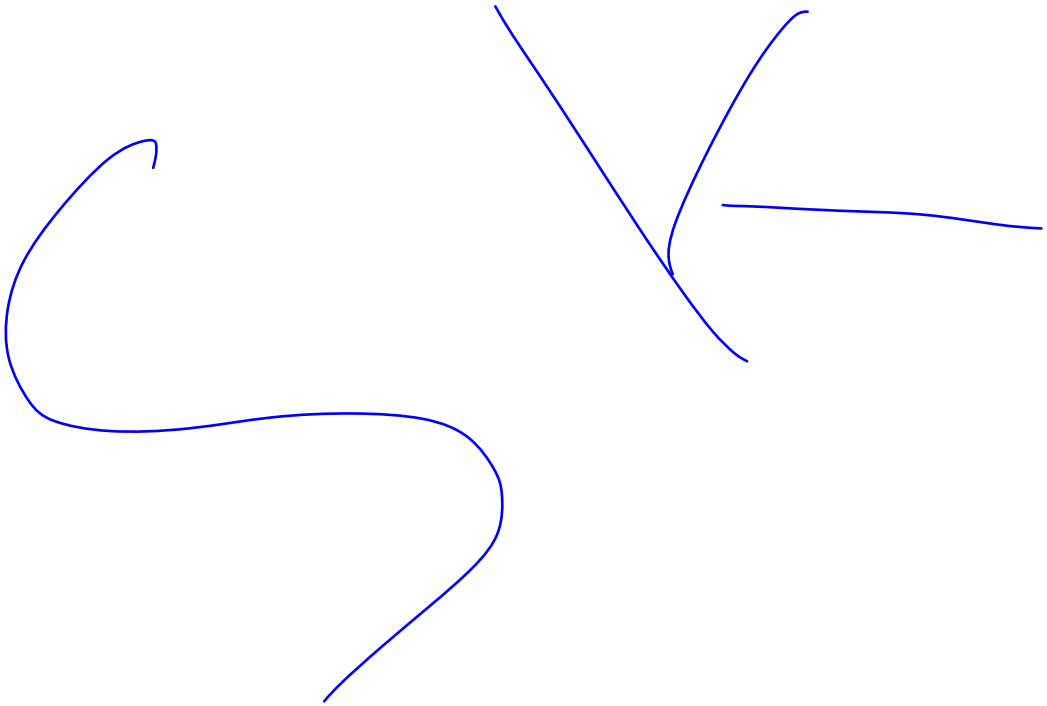
X	1	2	3	4	...
Pr	p	qp	q^2p	q^3p	...

$$\mu_X = \frac{1}{p} \text{ and } \sigma_X = \sqrt{\frac{q}{p^2}}$$

In general,

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

Example: Products produced by a machine have a 3% defective rate. What is the probability that the first defective occurs in the fifth item inspected?



Example: A shooter normally hits the target 70% of the time.

(a) Find the probability that her first hit is on the second shot.

(b) Find the mean and standard deviation.