Introduction to Statistical Analysis

Stat Bootcamp
Session 2

Sema Barlas

Discrete Probability Distributions

Statistics

Person #	1	2	3	4	5	6	7	8	9	10
X: Clicked?	N	Y	Y	Y	N	N	Y	Y	N	N
P	0	.5	.667	.75	.6	.5	.571	.625	.556	.5

$$P(X) = \frac{N(X{=}Y)}{N(X{=}Y \text{ } or \text{ } X{=}N)} = \frac{5}{10} = .5$$

Probability, p is given.

Whether a person would click the ad -> Bernoulli Trial.

Sample space: Yes and No (success and failure)

Probability, p and N are given.

Whether at least two people out of 10 would click the ad – Binomial trial.

X: # of people clicking

Probability

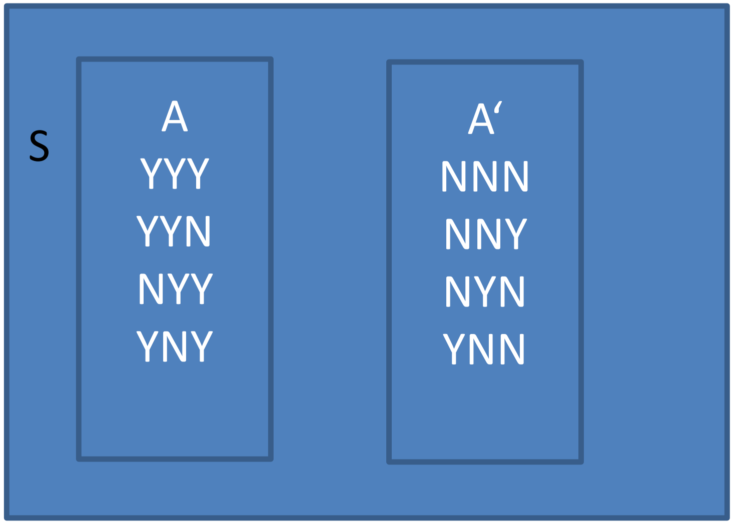
- You have 3 unique exposures to an ad in one hour. What is the probability for that at least 2 exposures are clicked - > $p(x \geq 2)$.
- Outcome for a single experiment: 2, Replication: 3
- Total number of outcomes in sample space: 2^3 = 8

Sample Space	YYY	YYN	YNY	YNN	NYY	NNY	NNY	NNN
X	3	2	2	1	2	1	1	0
p	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8

- $p(x \geq 2) = p(x=3) + p(x=2) = \frac{1}{8} + \frac{3}{8} = \frac{4}{8} = 0.5$
- $\sum_{i=1}^{8} P(O_i) = 1$

Complement

- Complement of event A, A', is the set of all outcomes that are not in A.
- A: at least two clicks {YYY, YYN, NYY, YNY}
- A‘: {NNN, NNY, NYN, YNN}



S

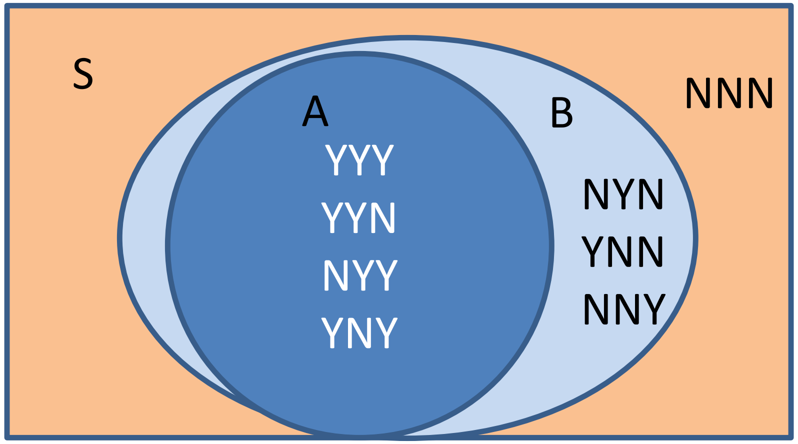
A
YYY
YYN
NYY
YNY

A‘
NNN
NNY
NYN
YNN

P(A) ≥ 0
P(A) ≤ 1
P(S) = 1
P(A) + P(A‘) = 1
P(A) = 1- P(A‘)
P(A') = 1- P(A)

Union and Intersection

- Union: A or B –> A U B
- Intersection: A and B –> A ∩ B
- A: at least two clicks {YYY, YYN, NYY, YNY}
- B: at least one click {YYY, YYN, NYY, YNY, NYN, YNN, NNY}
- A ∩ B: {YYY, YYN, NYY, YNY}
- A U B : {YYY, YYN, NYY, YNY, NYN, YNN, NNY}
- Events are disjoint or independent if A ∩ B = Ø –> P(A ∩ B) = 0.



P(A U B) = P(A) + P(B) – P(A ∩ B)

Application

Books	C1	C2	C3	C4
A	1	1	0	0
B	0	1	1	0
C	1	1	0	1
D	1	0	1	1

Jaccard similarity between customers = C1 ∩C2 / C1 U C2 = 2/4

Permutation

- There are 3 (n) web pages on your web site and visitors can access from one page to all other pages. Visitors usually select 2 pages (k). How many ways are there to select the 2 pages?
- 12, 13, 21, 23, 31, 32 (ordered subsets)
- 3*2
- $P_{k,n} = \frac{n!}{(n-k)!} = \frac{3!}{(3-2)!} = \frac{3*2*1}{1} = 6$
- $P_{2,4} = \frac{4!}{(4-2)!} = \frac{4*3*2*1}{2} = 12$

Combination

- There are 3 (n) web pages on your web site and visitors can access from one page to all other pages. Visitors usually select 2 pages (k). Which 2 pages are selected?
- 21, 13, 23 (unordered subsets)
- $\binom{n}{k}$ = $\frac{P_{k,n}}{k!}$ = $\frac{n!}{k!(n-k)!}$ = $\frac{3*2*1}{2*1*1}$ = 3

Conditional Probability

	B - faulty	B’ – not faulty
Line A	2	6
Line A’	1	9

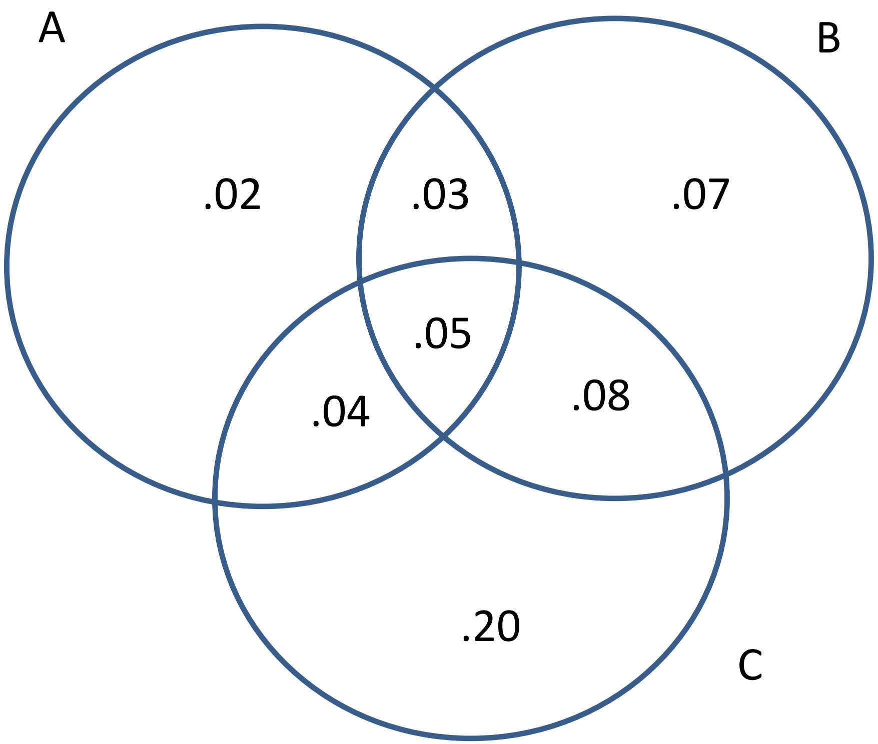
P(A) = $\frac{8}{18}$ = 0.44

P(A|B) = $\frac{2}{3}$ = $\frac{\frac{2}{18}}{\frac{3}{18}}$ = $\frac{P(A \cap B)}{P(B)}$

Reading Habits

A: Art, B: Books, C: Cinema

Read Regularly	A	B	C	A ∩ B	A ∩ C	B ∩ C	A ∩ B ∩ C
P	.14	.23	.37	.08	.09	.13	.05



$$P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{.08}{.23} = .348$$

$$P(A|BUC) = \frac{P(A\cap(BUC))}{P(BUC)} = \frac{.04+.05+.03}{.47} = .225$$

P(A|reads at least one) = P(A|A U B U C)

$$= \frac{P(A\cap(A\,U\,B\,U\,C))}{(A\,U\,B\,U\,C)}$$

$$= \frac{P(A)}{(A\,U\,B\,U\,C)} = \frac{.14}{.49} = .286$$

$$P(AUB|C) = \frac{P((AUB)\cap C)}{P(C)} = \frac{.04+.05+.08}{.37} = .459$$

Multiplication Rule for P(A∩B)

- $P(A \cap B) = P(A|B) * P(B)$

Player Brand	Market Share	Repair Rate
A	50%	25%
B	30%	20%
C	20%	10%

- Probability that a consumer bought Brand A that will need repair?
- Probability that customer has a player that will need repair?
- Given that player needs repair, what is the probability that it is brand A? Brand B? Brand C?

Independence

- If two events A and B are independent:
- $P(A|B) = P(A)$
- $P(A \cap B) = P(A) * P(B)$

A	B	AB	O
.40	.11	.04	.45

- What is the probability that blood phenotypes of two randomly selected individuals match?

Binomial Probability Distribution

- The experiment consists of a sequence of n smaller experiments called trials, where n is fixed in advance of the experiment.
- Each trial can result in one of the same two possible outcomes, success (S) or Failure (F).
- The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
- The probability of success P(S) is constant from trial to trial; we denote this probability by p.

- Examples: The number of heads when one flips a coin 10 times. Number of customers who pay with credit card among 10 customer who visit the store.

- $b(x; n, p)$
- $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$

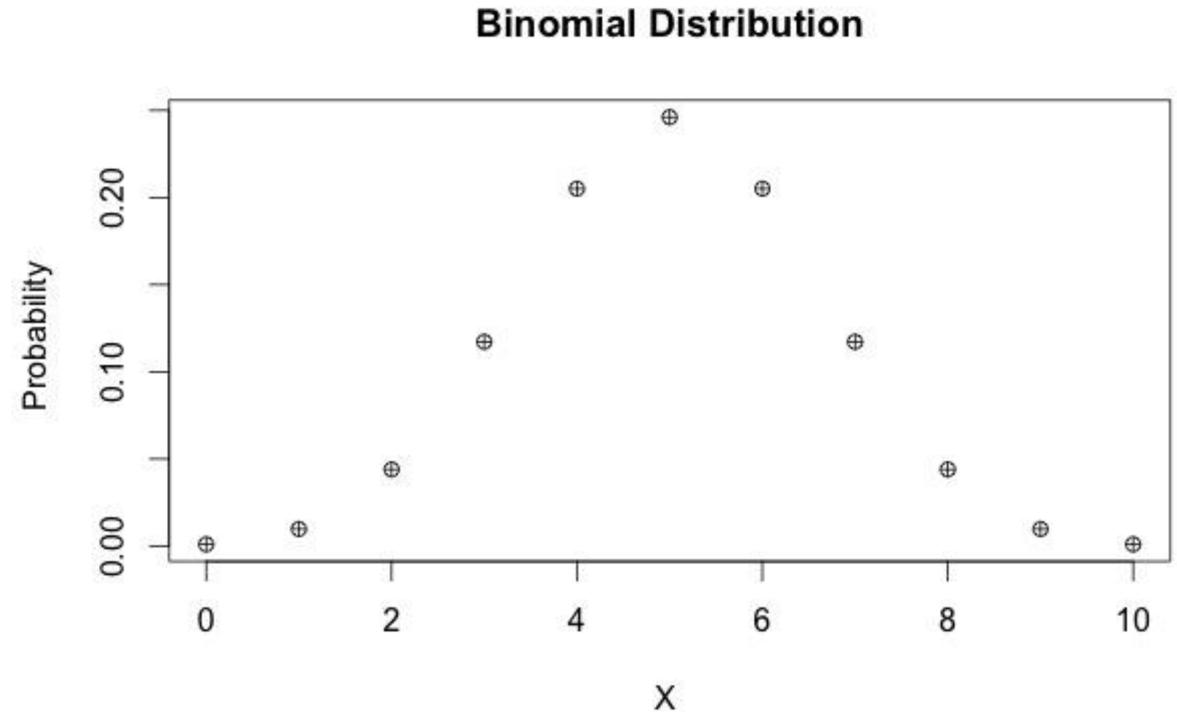
Example

- 20% of customers click your ad.
- Select random 5 people
- X: # of customers who click your ad.
- What is the probability that at most 3 customers click your ad?
- P(X=3) = b(3; 5, .20) = $\binom{5}{3}.20^3\ .80^2$= .0512
- P(X=2) = $\binom{5}{2}.20^2\ .80^3$= 0.2048
- P(X=1) = $\binom{5}{1}.20^1\ .80^4 = 0.4096$
- P(X=0) = $.80^5$=0.32768
- Answer: 0.99328

- Mean = E(X) = np = 5*.20 = 1
- Variance(X) = np(1-p)

Flipping a Fair Coin 10 Times

0	1	2	3	4	5	6	7	8	9	10
.0001	.001	.044	.117	.205	.246	.205	.117	.044	.001	.0001

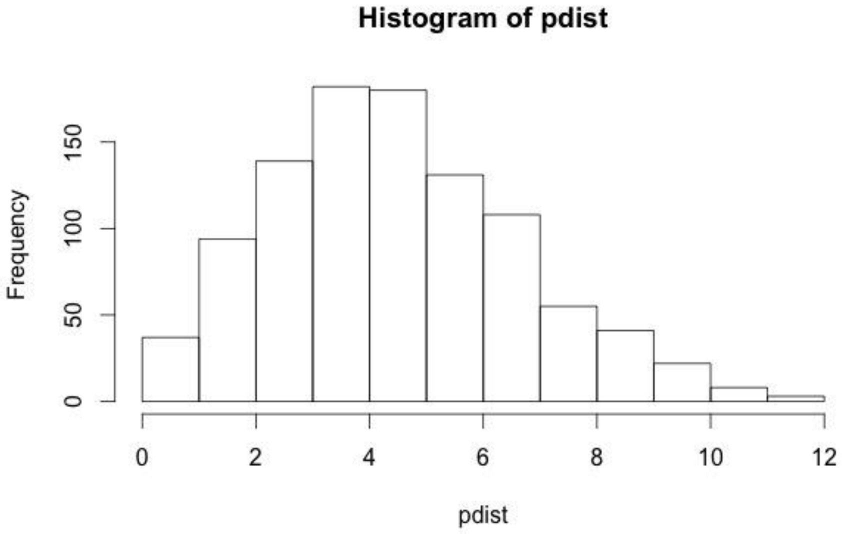


The Poisson Distribution

- $p(x; \mu) = \frac{e^{-\mu}\mu^{x}}{x!}, \mu > 0 \; and \; x \geq 0$
- # of visitors to a web page in an hour, µ = 4.5
- What is the probability that 5 people visits the webpage in an hour?
- $p(x = 5)$ = $\frac{e^{-4.5}4.5^{x}}{5!}$ =0.1708
- When n -> ∞, $binomial\ approaches\ poisson$. If n>50 and µ=np <5 we may use Poisson rather than binomial.
- $Mean = variance = \mu$

Example

- An article in the Los Angeles Times (Dec 3, 1993) repots that 1 in 200 people carry the defective gene that causes inherited colon cancer. In a sample of 1000 individuals, what is the approximate distribution of the number who carry the gene? What is p(5 ≤ x ≤ 8)?



Pdist=rpois(1000,5)

$p(x = 5) = \frac{e^{-5}5^5}{5!}$ = 0.175

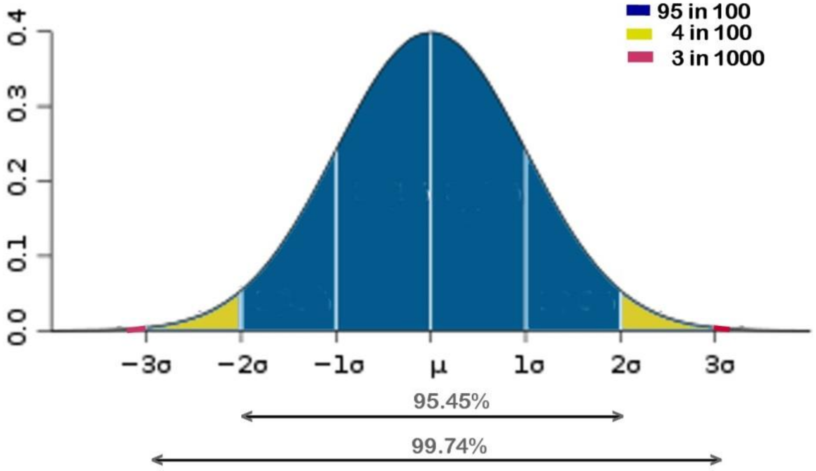
$p(x = 6) = \frac{e^{-5}5^6}{6!}$ = 0.146

$p(x = 7) = \frac{e^{-5}5^7}{7!}$ = 0.104

$p(x = 8) = \frac{e^{-5}5^8}{8!}$ = 0.065

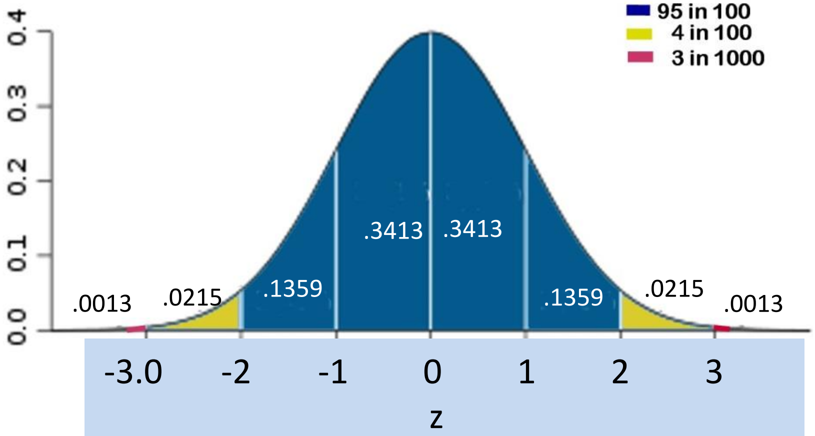
P(5 ≤ x ≤ 8) = 0.491

Continues Distributions
Normal Distribution



	Population Parameters	Sample Statistics
Mean	μ	$\bar{x}$
Variance	σ^2	Var
Standard Deviation	σ	s

Standard Normal Distribution



Z= $\frac{x-\mu}{\sigma}$ z= $\frac{x-\bar{x}}{s}$

Hypotheses

- Non-Directional Hypotheses
  - H0: $\mu = 100$
  - H1: $\mu \neq 100$
- Directional Hypotheses
  - H0: $\mu \leq 100$
  - H1: $\mu > 100$

Decision Making

	True State	
Decision	H0	H1
H0	Confidence	Type II mistake β
H1	Type I mistake α	Power

$\mu= 100, \sigma=10, \alpha=0.05$

Child	Seconds of Concentration	z	p
1	75		
2	81		
3	89		
4	99		
5	115		
6	127		
7	138		
8	139		
9	142		
10	148		

H0: Child comes from the distribution with $\mu=100$ and $\sigma=10$.
HA: Child does not comes from the distribution with $\mu=100$ and $\sigma=10$.

μ= 100 and σ=10

Child	Seconds of Concentration	z	p	Decision	Error Type
1	75	-2.50	0.006	Reject Null	Type I
2	81	-1.90	0.029	Reject Null	Type I
3	89	-1.10	0.136	Retain Null	Type II
4	99	-0.10	0.460	Retain Null	Type II
5	115	1.50	0.067	Retain Null	Type II
6	127	2.70	0.004	Reject Null	Type I
7	138	3.80	< 0.001	Reject Null	Type I
8	139	3.90	<0.001	Reject Null	Type I
9	142	4.20	< 0.001	Reject Null	Type I
10	148	4.80	<0.001	Reject Null	Type I

A test with $\bar{x}$=54.1 and s=13.41

- Top 10% will get an A. So, what is the cut-off point, assuming that the scores are normally distributed.
- Z = $\frac{x - 54.1}{13.41}$, x= 54.1 + 13.41*z, x=$\bar{x}$ + σ*z
- Z=1.28
- x=71.26

A test with $\bar{x}$=54.1 and s=13.41

- What proportion of students would have scores > 65?
- $z = \frac{65 - 54.1}{13.41}$ = 0.81
- P (z<0.81) = 0.791
- P(z>0.81) = 1-0.791 = 0.209

A test with $\bar{x}$=54.1 and s=13.41

- Less than 30?
- Between 45 and 85

Sampling Distribution

- Central Limit Theorem: Distribution of means approaches normal even when the underlying population is not normal.
- $\mu_{\bar{x}}$ = μ
- Standard error of the mean, $\sigma_{\bar{x}}$ = $\frac{\sigma}{\sqrt{n}}$