MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

Review

- Last time, finished Chapter 6. We learned basic event probability definition and notations.
- "union" $\cup$ corresponds to "or", "intersect" $\cap$ corresponds to "and".
- The additive and multiplicative rules.
- The conditional probability. (Bayes Theorem)

- Today we start the next topic on probability distributions in Chapter 7.

Chapter7 Probability Distributions

We have learned the probability rules on events.
Now we want to focus on random variables.

- A random variable: A numerical summary of the result from an experiment.
- Example: Toxicity study on 5 mice {A,B,C,D,E}.

Random variable X=number of mice died.

Possible outcomes become X=0,1,2,3,4,5

- Probability distribution: P(X=0), P(X=1),…,P(X=5)

Chapter7 Probability Distributions

We will cover three common probability distributions in this Chapter.

- Binomial distribution.
- Poisson distribution.
- Normal (also called Gaussian) distribution.
- The first two are for discrete random variables, while the last one is for a continuous random variable.

1. Binomial Distribution

The number of ‘successes’ in n tries follows a Binomial distribution. The characteristics:

1. The total number of trials is fixed (n).
2. Each trial has a binary outcome: success/failure.
3. The outcome of each trial is independent of all others.
4. Probability of ‘success’ is constant for all trails.

Example: Randomly poll 100 persons. Then *the number of females in the sample* ~ Binomial distribution.

1. Binomial Distribution

Example: Toss a coin two times and record results.

Outcomes: HH, HT, TH, TT. (H=head, T=Tail)

Random variable (R.V.) X ~ the number of heads.

So X=0 ↔ {TT}, X=1 ↔ {HT,TH}, X=2 ↔ {HH}.

For a fair coin, all outcomes are equally likely, hence P(X=0) = 1/4, P(X=1) =1/2 and P(X=2) = 1/4

1. Binomial Distribution (Coin tosses example)

R.V. X ~ the number of heads in two tosses.

If we have a cheater's coin where P(H)= 2/3, then due to independence,

$P(HH)=P(H)\cdot P(H)=(2/3)^2=4/9$, $P(HT)=(2/3)(1/3)=2/9$

$P(TH)=(1/3)(2/3)=2/9$, $P(TT)=(1/3)(1/3)=1/9$

Hence $P(X=0)=1/9$ $=(1/3)(1/3)$

$P(X=1)=2/9+2/9=4/9$ $=2(1/3)(2/3)$

and $P(X=2)=4/9$ $=(2/3)(2/3)$

1. Binomial Distribution

In general, X ~ Bin(n,p), then

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

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$, and $n! = n(n-1)\ldots(2)1$

For cheater's coin example, X ~ Bin(n=2,p=2/3),

$$P(X=0) = \binom{2}{0} \left(\frac{2}{3}\right)^0 \left(\frac{1}{3}\right)^2 = \frac{2!}{0!2!}(1)\left(\frac{1}{3}\right)^2 = \left(\frac{1}{3}\right)^2,$$

$$P(X=1) = \binom{2}{1} \left(\frac{2}{3}\right)^1 \left(\frac{1}{3}\right)^1 = \frac{2!}{1!1!}\left(\frac{2}{3}\right)\left(\frac{1}{3}\right) = \frac{4}{9}.$$

1. Binomial Distribution

Example: The incidence of chronic brochities among infants (≤1 year old) in the USA is 5%. In a random sample of 20 babies whose both parents have chronic bronchitis, 3 infants were found to have the disease.

How likely does this happen by chance?

Solution: Let X be the number of infants out of 20 having chronic bronchitis. If no relationship with parents' disease status, then X ~ Bin(n=20,p=0.05)

1. Binomial Distribution

Chronic brochities example solution (continued):

Let X be the number of infants out of 20 having chronic bronchitis. If no relationship with parents having the disease, then X ~ Bin(n= , p=)

$P(X \geq 3) = 1 - P(X=0) - P(X=1) - P(X=2)$

$$= 1 - \sum_{k=0}^{2} \binom{20}{k} (0.05)^k (0.95)^{n-k}$$

$$= 1 - (0.3585 + 0.3774 + 0.1887) = 0.0754$$

2. Poisson Distribution

Poisson RV reflects the count of events that are independently occurring uniformly in time or space.

There is no upper bound of the possible values that Poisson RV can take (in contrast to Binomial).

Formula for Poisson(λ) distribution

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

where λ reflects the mean (average).

Poisson is a limit (approximation) of Binomial

Binomial(n, p= λ/n) formula

$$P(X=k) = \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{n(n-1)\ldots(n-k+1)(n-k)\ldots(1)}{k!\ (n-k)\ldots(1)}\frac{\lambda^k}{n^k}\frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k}$$

When k and λ are fixed and $n \to \infty$, this becomes

$$P(X=k) = \frac{\lambda^k}{k!}\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = \frac{\lambda^k}{k!}e^{-\lambda}$$

2. Poisson Distribution

Example:

Migrating whales pass an observational point at the rate of 1.5 per hour.

Let X ~ number of whales sighted in one hour, then X ~ Poisson(λ=).

The probability of seeing no whale in one hour is

$$P(X=0) = \frac{1.5^0}{0!} e^{-1.5} = e^{-1.5} = 0.22$$

2. Poisson Distribution

Whale Example (continued):

The probability of seeing three or more whales in one hour is

$P(X \geq 3) = 1 - P(X=0) - P(X=1) - P(X=2)$

$$= 1 - \sum_{k=0}^{2} \frac{1.5^k}{k!} e^{-1.5}$$

$$= 1 - (0.22 + 0.33 + 0.25)$$

$$= 1 - 0.8 = 0.2$$

3. Normal Distribution

A normal RV is a *continuous* random variable, it can take any real number as its value.

A normal RV X ~ $N(\mu,\sigma^2)$, with two parameters:

μ = mean, σ = standard deviation (population)

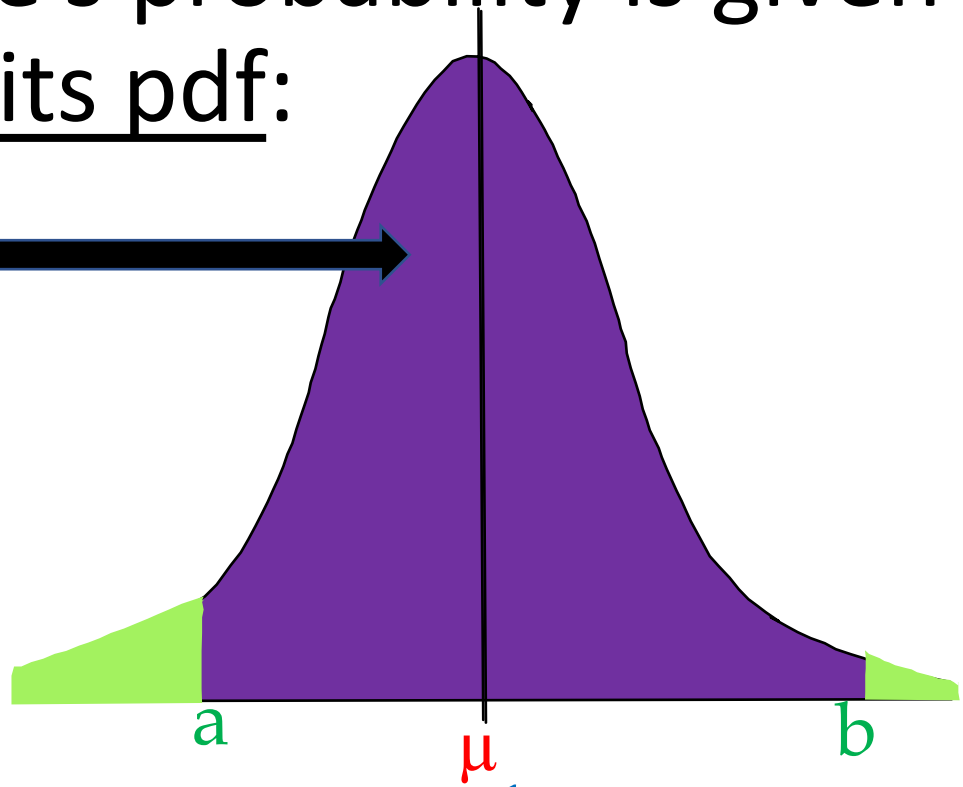
Normal probability density function(pdf):

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

3. Normal Distribution

Continuous random variable's probability is given by the area under curve of its pdf:

$$P(a \le X \le b) = \int_a^b f(x)dx$$



Normal pdf is bell-shaped $f(x)=\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

3. Calculation of Normal Probability

It is easier to find a *continuous* random variable's probability through its *cumulative distribution function* (cdf): $P(X \le x)$

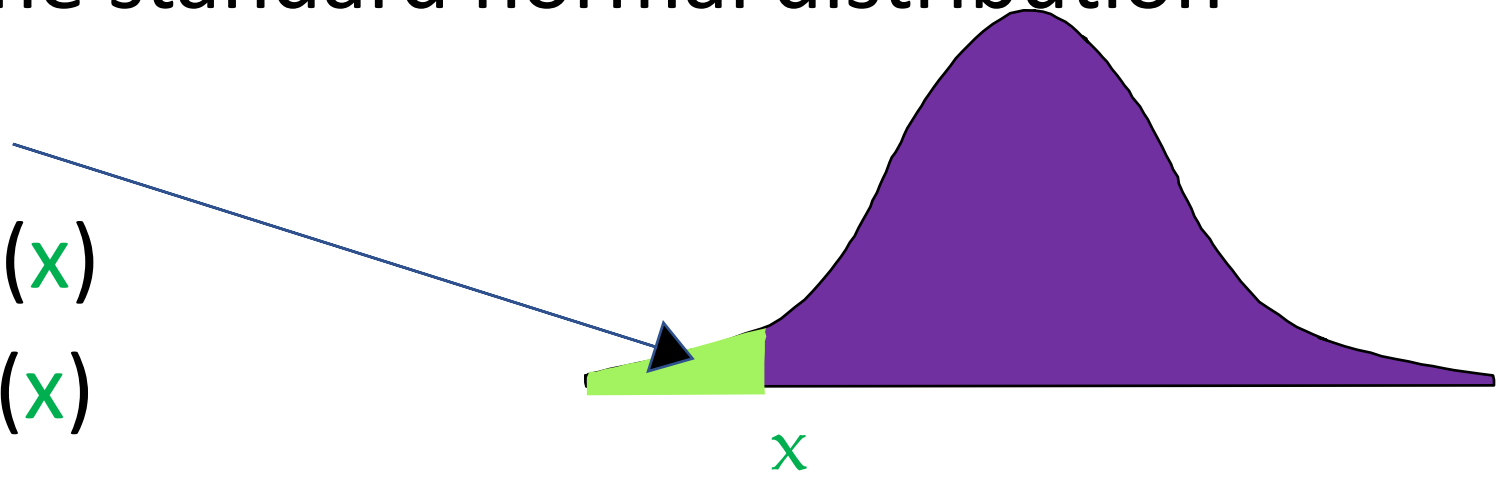
First, for $X \sim N(\mu, \sigma^2)$,

then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ the standard normal distribution

Denote $\Phi(x) = P(Z \le x)$

Hence $P(Z > x) = 1 - \Phi(x)$

$P(x < Z \le y) = \Phi(y) - \Phi(x)$



3. Calculation of Normal Probability

For $X \sim N(\mu,\sigma^2)$, then $Z=\frac{X-\mu}{\sigma} \sim N(0,1)$

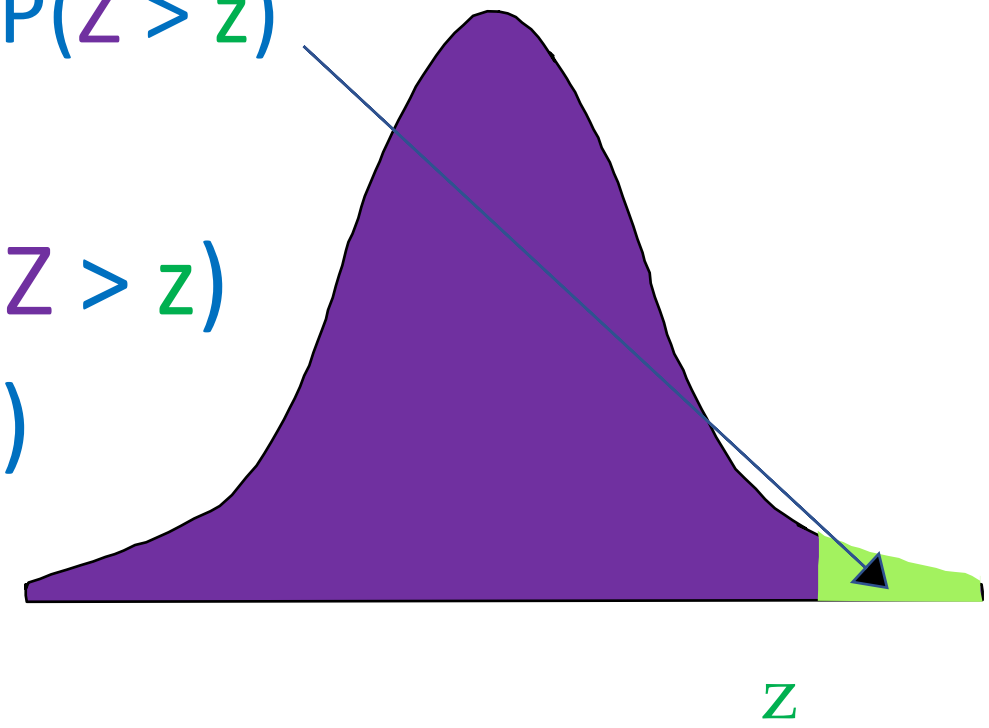
How to find $\Phi(x)=P(Z \le x)$?

In textbook, Table A.3 gives $P(Z > z)$

Can find $\Phi(x)=P(Z \le x)=1-P(Z > z)$

Or directly find $P(x< Z \le y)$

$= P(Z > x) - P(Z > y)$



3. Normal Distribution

Example: The heights of Northeastern University senior male (NUSM) students are normally distributed with mean 70 inches and standard deviation 2.6 inches.

What proportion of NUSMs are over 74 inches tall?

Solution: Let X be the height of on NUSM, then X~N(μ= 70, $\sigma^2 = 2.6^2$). How to calculate P(X≥74)?

3. Normal Distribution

Height Example continued:

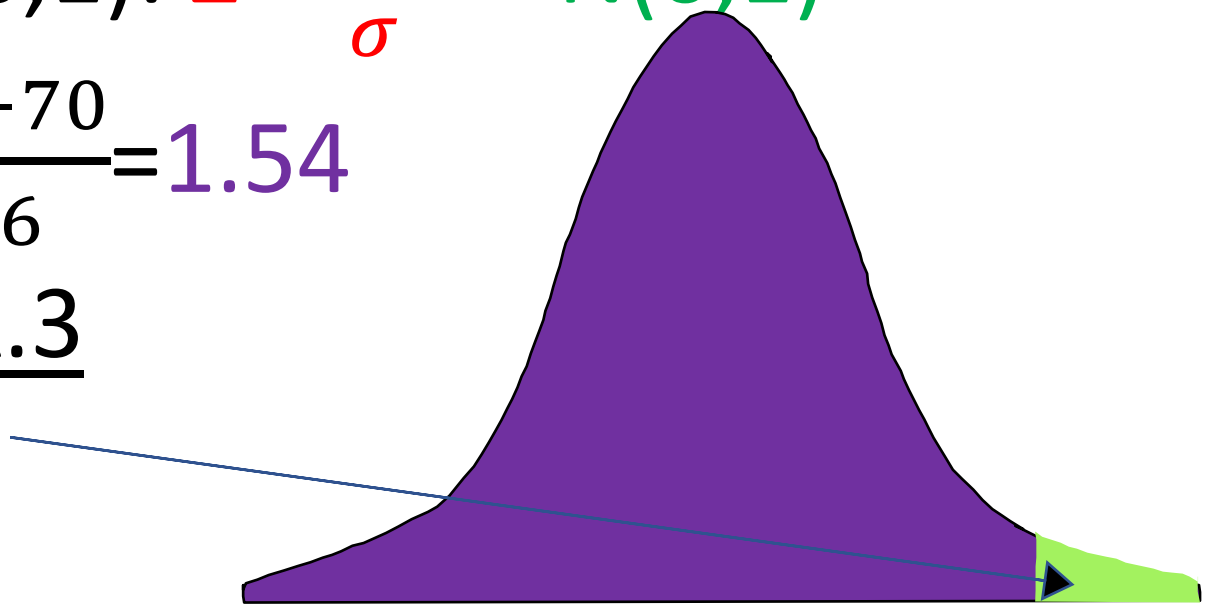
X~N(μ= 70, $\sigma^2 = 2.6^2$). How to calculate P(X≥74)?

Step 1. Rescale to N(0,1): $Z=\frac{X-\mu}{\sigma}$ ~ N(0,1)

X≥74 → $Z \geq \frac{74-\mu}{\sigma}=\frac{74-70}{2.6}=1.54$

Step 2. Check Table A.3

P(X≥74)=P(Z ≥1.54)



1.54

3. Normal Distribution

Step 2.

Check Table A.3

P(X≥74)=P(Z ≥1.54)

= 0.062

TABLE A.3

Areas in the upper tail of the standard normal distribution

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.500	0.496	0.492	0.488	0.484	0.480	0.476	0.472	0.468	0.464
0.1	0.460	0.456	0.452	0.448	0.444	0.440	0.436	0.433	0.429	0.425
0.2	0.421	0.417	0.413	0.409	0.405	0.401	0.397	0.394	0.390	0.386
0.3	0.382	0.378	0.374	0.371	0.367	0.363	0.359	0.356	0.352	0.348
0.4	0.345	0.341	0.337	0.334	0.330	0.326	0.323	0.319	0.316	0.312
0.5	0.309	0.305	0.302	0.298	0.295	0.291	0.288	0.284	0.281	0.278
0.6	0.274	0.271	0.268	0.264	0.261	0.258	0.255	0.251	0.248	0.245
0.7	0.242	0.239	0.236	0.233	0.230	0.227	0.224	0.221	0.218	0.215
0.8	0.212	0.209	0.206	0.203	0.200	0.198	0.195	0.192	0.189	0.187
0.9	0.184	0.181	0.179	0.176	0.174	0.171	0.169	0.166	0.164	0.161
1.0	0.159	0.156	0.154	0.152	0.149	0.147	0.145	0.142	0.140	0.138
1.1	0.136	0.133	0.131	0.129	0.127	0.125	0.123	0.121	0.119	0.117
1.2	0.115	0.113	0.111	0.109	0.107	0.106	0.104	0.102	0.100	0.099
1.3	0.097	0.095	0.093	0.092	0.090	0.089	0.087	0.085	0.084	0.082
1.4	0.081	0.079	0.078	0.076	0.075	0.074	0.072	0.071	0.069	0.068
1.5	0.067	0.066	0.064	0.063	0.062	0.061	0.059	0.058	0.057	0.056
1.6	0.055	0.054	0.053	0.052	0.051	0.049	0.048	0.047	0.046	0.046
1.7	0.045	0.044	0.043	0.042	0.041	0.040	0.039	0.038	0.038	0.037

Three Probability Distributions covered

- (1) Know when to use which distribution:

  Binomial, discrete, bounded

  Poisson, discrete, unbounded

  Normal, continuous, bell-shaped

- (2) Calculate probability from the distributions:

  a. discrete formula plug-in (or Table A1, A2)

  b. N(μ,σ^2): (i) Scale $\frac{X-\mu}{\sigma}$~N(0,1), (ii) Table A.3

Use R to calculate Probability

- R has coded the calculation of cdf: $P(X \le x)$ in form of p__(), with name of distribution in ___.
- Instead of finding normal probability $P(Z > z)$ from Table A.3, we use R to find $\Phi(z)=P(Z \le z)$, with pnorm(z)

- In the height example before,

$P(X \ge 74)=P(Z \ge 1.54)=$ 1 - pnorm(1.54)

Use R to calculate normal Probability

- Alternatively, we can directly use

pnorm(x, mean= , sd=) to find the cdf of N(μ,σ^2).

- In the height example before,

P(X≥74)= 1- pnorm(74, mean=70, sd=2.6)

While it is much easier to do it this way, you should know that $Z=\frac{X-\mu}{\sigma}$ ~ N(0,1)

Use R to calculate Probability

- Binomial cdf $P(X \le x)$ = pbinom(x, size= , prob=)

  Poisson cdf $P(X \le x)$ = ppois(x, lambda=)

- In previous examples: $X \sim Bin(n= 20, p=0.05)$

  $P(X \ge 3) = 1 - P(X < 3) = 1 - P(X \le 2)$ (discrete)

  = 1 - pbinom(2, size= 20 , prob= 0.05)

  $X \sim Poisson(\lambda= 1.5)$, possible value X=0, 1, 2, …

  $P(X=0) = P(X \le 0)$ = ppois(0, lambda= 1.5)

  $P(X \ge 3)$ = 1 - ppois(2, lambda= 1.5)

Other Probability distributions seen later

- Chi-square distribution (degrees of freedom =k):

$Z_1,\ldots, Z_k$ i.i.d.~ N(0,1), then X=$Z_1^2 + \cdots + Z_k^2 \sim \chi^2_{df=k}$

In R, the cdf P(X≤x) is given by pchisq(x, df=)

- T-distribution (with k degrees of freedom)

$Z_0,\ldots, Z_k$ i.i.d.~ N(0,1), then X=$\dfrac{Z_0}{\sqrt{(Z_1^2+\cdots+Z_k^2)/k}} \sim t_{df=k}$

In R, the cdf P(X≤x) is given by pt(x, df=)

Other Probability distributions seen later

- F-distribution

$Z_1,\ldots, Z_m, Y_1,\ldots, Y_n$ i.i.d. ~ N(0,1), then

$$X=\frac{(Z_1^2+\cdots+Z_m^2)/m}{(Y_1^2+\cdots+Y_n^2)/n} \sim F_{df_1=m,df_2=n}$$

In R, the cdf P(X≤x) is given by pf(x, df1= , df2=)

- The Chi-square, t and F distributions are utility distributions. Just need to know their definition, and R command for calculation

Theoretical Mean and Variance of R.V.s

- (center) Mean E(X)= $\begin{cases} \sum x_i P(x_i) & discrete\ r.v. \\ \int xp(x)dx & continuous\ r.v. \end{cases}$

  (spread) Variance Var(X) = E[X - E(X)]^2

Distribution	**parameters**	**Mean**	**Variance**
Binomial	n,p	np	np(1-p)
Poisson	λ	λ	λ
Normal	μ,σ^2	μ	σ^2

Theoretical Mean and Variance of R.V.s

Distribution	parameters	Mean	Variance
Chi-square	df = k	k	2k
t	df = k	0	$\begin{cases} \infty, & k \le 2 \\ \dfrac{k}{k-2}, & k > 2 \end{cases}$
F	$df_1 = k_1$, $df_2 = k_2$	$\begin{cases} \infty, & k_2 \le 2 \\ \dfrac{k_2}{k_2-2}, & k_2 > 2 \end{cases}$	$\begin{cases} \infty, & k_2 \le 4 \\ \dfrac{2(k_2)^2\,(k_1+k_2-2)}{k_1(k_2-2)^2\,(k_2-4)}, & k_2 > 4 \end{cases}$

- The Chi-square, t and F distributions are utility distributions. Just need to know their definition, and R command for cdf calculation

Summary

- We introduced three probability distributions

- Should know when to use which, and how to calculate the probability (from cdf or Table A.?)
- Mean and Variance

- Next time, more on properties of Mean and Variance, and more on approximations of probability distributions