# MATH 7343 Applied Statistics

Prof. (Aidong) Adam Ding

Northeastern University

# Review

- Last time, we learned about probability distributions, particularly Binomial, Poisson and normal distributions. Briefly mentioned chi-square, t and F distributions. We also started on the Mean and Variance.
- Today we will finish the topic on probability.

- Project teams are assigned. Preliminary proposal due on March10

# 1. Properties of Mean and Variance

- Property of the Mean: R.V.s X and Y, constants a and b.

$$E(aX+bY) = a\,E(X) + b\,E(Y)$$

- Example:

X ~monthly income of husband in a two-incomes family

Y ~monthly income of wife in a two-incomes family

If we know E(X) and E(Y), what is the mean annual income of a two-incomes family?

Solution: E( ) = +

# 1. Properties of Mean and Variance

Let X and Y be R.V.s, a and b be constants.

- <u>Property of the Mean</u>:

  $E(aX+bY) = a\,E(X) + b\,E(Y)$;

- <u>Property of the Variance</u>:

  $Var(aX) = a^2\,Var(X)$ always,

  $Var(X+Y) = Var(X) + Var(Y)$ ***<u>IF X and Y are independent</u>***

# 1. Properties of Mean and Variance

- Example: Support for legal abortion in males is 55%;

Support for legal abortion in females is 65%. We sample 150 males and 162 females in a survey about abortion attitude.

$X$ ~ number of legal abortion supporters in the sample

What is the mean and variance of $X$?

- Solution: Let $X_m$ ~ # of male supporters in the sample, and

  $X_f$ ~ # of female supporters in the sample;

  so $X = X_m + X_f$

  $X_m \sim Bin(n=150, p=0.55)$, $X_f \sim Bin(n=162, p=0.65)$

# 1. Properties of Mean and Variance

- Survey Example Solution (continued):

  $X$ ~ # of legal abortion supporters in the sample.

  $X = X_m + X_f$

  $X_m \sim Bin(n=150, p=0.55)$, $X_f \sim Bin(n=162, p=0.65)$

Hence $E(X) = E(X_m + X_f) = E(X_m) + E(X_f) = \ \ + \ = 187.8$

$Var(X) = Var(X_m) + Var(X_f) = + = 73.98$

$Sd(X) = \sqrt{73.98} = 8.60$

Notice that the variance calculation used the independence between $X_m$ and $X_f$.
(The mean calculation do not need this assumption.)

# 1. Properties of Mean and Variance

- Survey Example Discussion:

  $X$ ~ # of legal abortion supporters in the sample.

For our solution, we used that $X = X_m + X_f$ with

  $X_m \sim \text{Bin}(n=150,\ p=0.55)$, $X_f \sim \text{Bin}(n=162,\ p=0.65)$

Another way to look at this: In population, 60% support.
Sampled 312 persons, so # of supporters ~$\text{Bin}(n=312,\ p=0.60)$.

Can we just find mean and variance of $\text{Bin}(n=312,\ p=0.60)$?

When should we use $\text{Bin}(n=150,\ p=0.55)+\text{Bin}(n=162,\ p=0.65)$

and when should we use $\text{Bin}(n=312,\ p=0.60)$?

# 2. Properties of Normal Random Variables

- Property: Linear combination of normal R.V.s is still normal.
- Say, $X \sim N(\mu_X, \sigma_X^2)$ is independent of $Y \sim N(\mu_Y, \sigma_Y^2)$, then what is the distribution of aX+bY?

Solution:

Using the mean property $E(aX+bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y$

Using the variance property

$$Var(aX+bY) = a^2\,Var(X) + b^2\,Var(Y) = a^2\sigma_X^2 + b^2\sigma_Y^2$$

Then use the normal R.V.s property

$$aX+bY \sim N(a\mu_X + b\mu_Y,\ \ a^2\sigma_X^2 + b^2\sigma_Y^2)$$

# 2. Properties of Normal Random Variables

- Example: A nutrition diet is tried on mice for a week. Mice group A eats this nutrition diet; group B eats normal diet.

Let $X_A$ = average weight gain in group A ~ N($\mu_A, \sigma^2 = 1$)

$X_B$ = average weight gain in group B ~ N($\mu_B, \sigma^2 = 1$)

How do we check if the diet is working?

$E(X_A - X_B) = E(X_A) - E(X_B) = \mu_A - \mu_B$

$Var(X_A - X_B) = Var(X_A) + Var(-X_B) = 1 + (-1)^2\, 1 = 2$

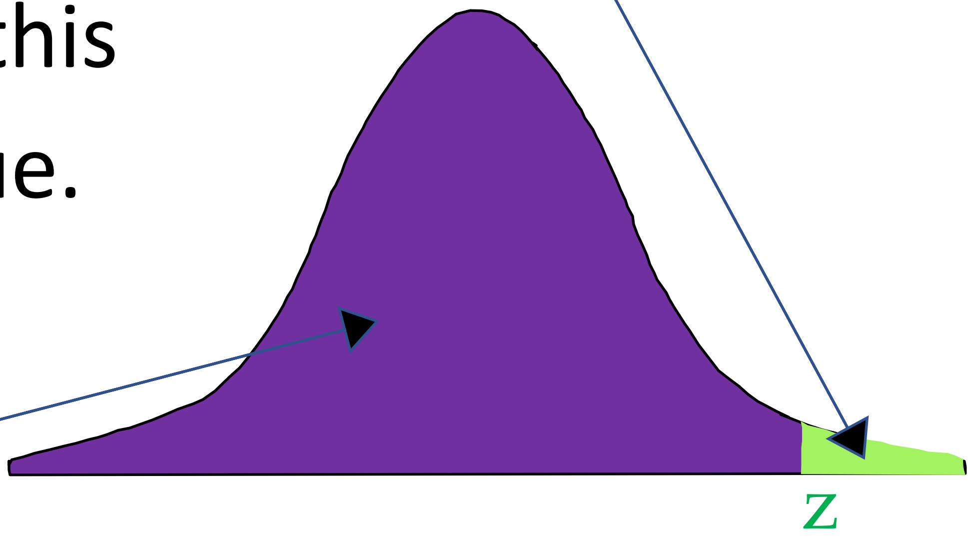
Hence $X_A - X_B \sim N(\mu_A - \mu_B, 2)$ . If diet has no effect, ~ N(0, 2)

Say, we observe $X_A - X_B = 20$, $P(N(0, 2) \geq 20) = ?$ 0.000

Can this happen by chance?

# 3. Find Quantiles of Normal Distribution

Before, we used Table A.3 in textbook to find p=P(Z > z) for N(0,1). We can also reverse this process to find z for given p value.




Or use R:

given P(Z ≤z) find corresponding quantile z value.

# 3. Find Quantiles of Normal Distribution

Example: SAT math scores are normally distributed with mean 550 and standard deviation 100. What is the score needed to be in the 80-th percentile?

Solution: Table A.3 look for $p=P(Z > z)=1-0.8=0.2$, we get $z=0.84$.

Since $Z=\frac{X-\mu}{\sigma}$, so $X=\sigma Z+\mu$

Score needed is $100(0.84)+550 = 634$.

# 3. Find Quantiles of Normal Distribution

$P(Z > z)=0.2$

$z = 0.84$

**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 |

# 3. Find Quantiles of Normal Distribution

Example: SAT math scores are normally distributed with mean 550 and standard deviation 100. What is the score needed to be in the 80-th percentile?

Solution: Table A.3 look for p=P(Z > z)=1-0.8=0.2, we get z=0.84. Since $Z=\frac{X-\mu}{\sigma}$, so $X=\sigma Z+\mu$

Score needed is 100(0.84)+550 = 634.

Easier using R: qnorm(0.8)*100+ 550

or qnorm(0.8, mean=550, sd=100)

# 4. Approximations of Binomial Distribution

(1) Poisson(λ) approximates Bin(n, p)

when np=λ and $n \to \infty$

Example: X~ Bin(n=1000, p=0.002)

Which Poisson best approximate it?

Match mean np=(1000)(0.002) = 2 = λ.

So Poisson(λ=2) best approximates X.

So P(X=0) ≈ $e^{-2}$

Notice that variance not exact match here.

Var = np(1-p) = 1000(0.002)(1-0.002) = 2(0.998) ≈λ but ≠λ

# 4. Approximations of Binomial Distribution

(2) Normal approximation of Bin(n, p) when $n \to \infty$,

and $p$ is not too small nor too large

Example: X~ Bin(n=100, p=0.1), better approximated with normal rather than Poisson distribution.

Which N(μ,$\sigma^2$) best approximate it?

Match mean np =(100)(0.1) = 10 = μ.

Variance np(1-p) = 100(0.1)(1-0.1) = 9 $= \sigma^2$.

So X ≈ N(μ=10, $\sigma^2 = 9$)

# 4. Approximations of Binomial Distribution

## (2) Normal approximation of Bin(n, p)

Example(continued): X~Bin(n=100, p=0.1) ≈N(μ=10, $\sigma^2 = 9$)

So P(9≤X≤11) with continuity correction

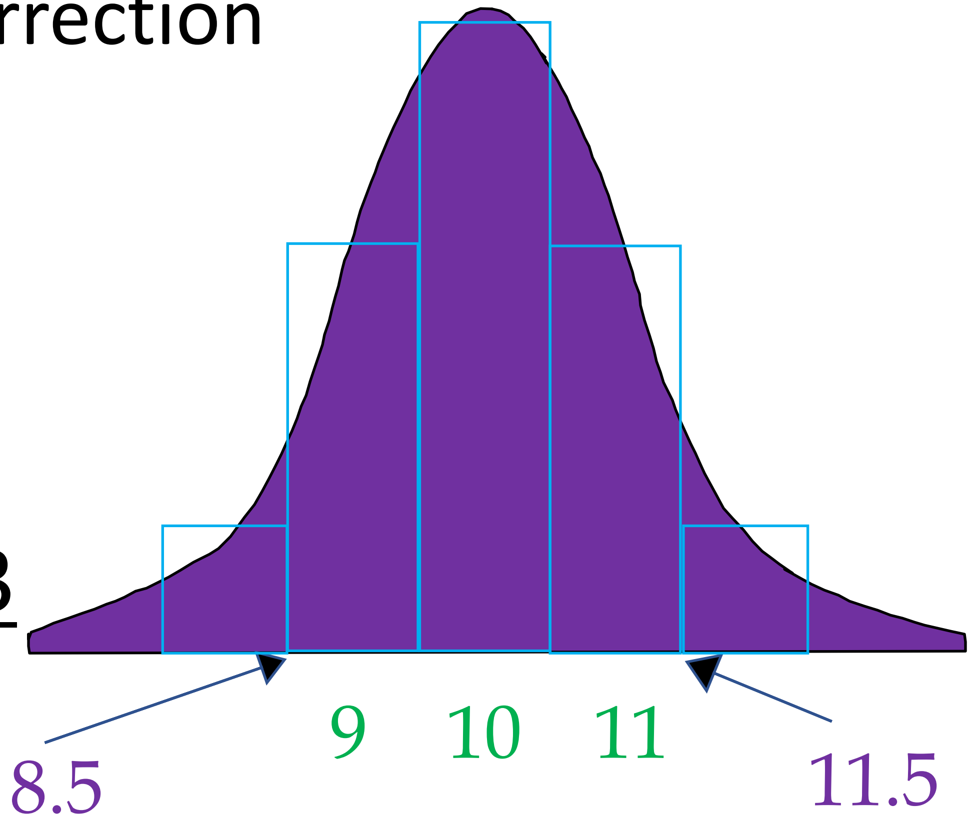
= P(8.5<X<11.5)

= P($\frac{8.5-10}{\sqrt{9}}$<Z< $\frac{11.5-10}{\sqrt{9}}$)

= P(-0.5<Z<0.5)

= 1-2P(Z≥0.5) from Table A.3

= 1- 2(0.309) = 0.328



# 4. Approximations of Binomial Distribution

- Example: In a normal human, 60% of white blood cells (WBC) are neutrophils. In a WBC screen, 100 cells are counted. If <50 or >70 cells are neutrophils, a person is classified as abnormal. What proportion of normal person will be misclassified by this procedure? (Specificity of test)
- Solution: $X$~ number of neutrophils out of 100 cells

  ~Bin(n=100, p=0.6) for a normal person.

So P(misclassify as abnormal | normal person)

= $P(X<50 \text{ or } X>70)$

# 4. Approximations of Binomial Distribution

- WBC Example: X~ Bin(n=100, p=0.6) for a normal person.

P(X<50 or X>70) = 1 - $\sum_{k=50}^{70} \frac{100!}{k!(100-k)!}(0.6)^k(0.4)^{100-k}$

Hard to do by hand: involves 21 terms.

Normal approximation: np=60 = μ, np(1-p) = 24 = $\sigma^2$.

P(X<50 or X>70) = 1 – P(50≤X≤70) = 1 – P(49.5<X<70.5)

= 1 - P($\frac{49.5-60}{\sqrt{24}}$<Z< $\frac{70.5-60}{\sqrt{24}}$)

= 1 - P(-2.14<Z<2.14) = 2 P(Z≥2.14) = 2(0.016) = 0.032

Table A.3.

# 4. Approximations of Binomial Distribution

- Do we really need normal approximation for Binomial?

In R, normal approximation

```
pnorm((70.5-60)/sqrt(24)) – pnorm((49.5-60 )/sqrt(24))
```

Exact Binomial:

```
pbinom(70.5,size=100,prob=0.6) - pbinom(49.5,size=100,prob=0.6)
```

- With R, we do not need normal approximation for calculation.
- Why do we still teach this? You should know that, when n big, Binomial can be approximated by normal distribution. And you should know how to choose the distribution for best approximation: match mean and variance.

# 4. Approximations of Binomial Distribution

- R commands on Homework2 handout:
- We want to see how close the CDFs $P(X \le x)$ are for the Binomial approximations by Poisson and normal.

To do this, we calculate $P(X \le x)$ for x=0,1,…,20

Assign x values by x=(0:20), can calculate Bin(1000,0.01) CDF at those 21 values together by pbinom(x, size=1000, p=0.01)

Poisson approximation using $\lambda=1000(0.01)=10$, get CDF by

ppois(x, lambda=10); Normal approximation using $\mu=10$, $\sigma^2=1000(0.01)(1-0.01)=9.9$, get CDF by

pnorm((x+0.5-10)/sqrt(9.9))

# 4. Approximations of Binomial Distribution

- Homework2 handout: we then get a table

| | x | binprob | posprob | normprob |
|---|---|---|---|---|
| 1 | 0 | 0.00004 | 0.00005 | 0.00127 |
| 2 | 1 | 0.00048 | 0.00050 | 0.00345 |
| 3 | 2 | 0.00268 | 0.00277 | 0.00857 |
| 4 | 3 | 0.01007 | 0.01034 | 0.01942 |
| … | … | | | |
| 19 | 18 | 0.99310 | 0.99281 | 0.99655 |
| 20 | 19 | 0.99671 | 0.99655 | 0.99873 |
| 21 | 20 | 0.99850 | 0.99841 | 0.9995 |

We can see Poisson approximation is better than normal approximation here.

# 4. Approximations of Binomial Distribution

- R commands on Homework2 handout:
- To produce the table, we put all four variable in one data.frame probTable

  And display the table, rounding to 5 digits

  round(probTable, digits=5)

# 4. Approximations of Binomial Distribution

- R commands on Homework2 handout:
- To calculate CDF $P(X \leq x)$ we used

```
pbinom(x, size=1000, p=0.01)
```

What if we want to calculate the PDF $P(X=x)$ instead?

For the Binomial distribution, it is discrete and

$$P(X=x) = P(X \leq x) - P(X \leq x-1) = P(X \leq x+0.5) - P(X \leq x-0.5)$$

So we can calculate the PDF using

```
pbinom(x, size=1000, p=0.01)-pbinom(x-1, size=1000, p=0.01)
```

# 5. Properties for Mean and Variance of R.V.s

- R.V.s X and Y, constants a and b.

$E(aX+bY) = a\,E(X) + b\,E(Y)$ always,

$E(X-Y) = E(X) - E(Y)$ always,

$E(XY) \neq E(X)\,E(Y)$ usually,

$E(XY) = E(X)\,E(Y)$ *when X and Y are independent*

$Var(X+Y) \neq Var(X) + Var(Y)$ usually,

$Var(aX+bY) = a^2\,Var(X) + b^2\,Var(Y)$ *when X and Y are independent*

$Var(X-Y) = Var(X) + \cancel{(-1)^2}\,Var(Y)$ *when X and Y are independent*

# 5. Properties for Mean and Variance of R.V.s

- Example:

If the income of husbands has mean \$47,000 and variance $(\$20,000)^2$

the income of wives has mean \$44,000 and variance $(\$20,000)^2$

Then the incomes of married couples have

mean = 47,000 + 44,000 = \$91,000

variance = 2 $(\$20,000)^2$ if the incomes are independent between husband and wife

Since the independence is unlikely, in fact, we **can not find** the variance from the given information.

The mean difference between incomes of husband and wife is

47,000 - 44,000 = \$3,000

# Summary

- We finished probability review

You should know how to

(1) recognize when to use Binomial/Poisson

(2) calculate normal probability

(3) Calculate Binomial probability

  (a) use exact formulas

  (b) approximation by normal distribution

(4) Use R to find probability and quantile

(5) Properties of Mean and Variance

# Summary

- We finished probability review

- Next time, we use the probability theory to study how reliable the sample mean is as an estimator.

Homework 2 is due in One week

Get together with your project teams.