

# Basic statistics

April 24, 2019



# Data



# What is data?

- Data is set of numbers presenting values of a quantity
- Data is collected from real world
- From data we want to know about the quantity



# Population

- Population is the set of all values of the quantity in real world
- Ex: quantity = height of men in a city
- Population = the heights of all men in the city
- Size of population is large

# Sample

- Because Population is too large, we can not collect all data in Population
- Instead we collect data from Samples
- Samples = some values *chosen at random* from Population
- Ex: Choose some men in the city to measure their height in a random way: different locations, ages, incomes ...

# What is Statistics

- Statistics are the parameters can be calculated from Data
- Assuming that Data are samples from a random quantity
- Statistics gives us properties of the random quantity
- Can use Statistics to predict value or compare different data

# Mathematical view

- We will use Mathematical Probability to model Population, Samples and Statistics
- The Quantity = a Random Variable  $X$
- Population = Range of  $X$
- $X$  has the same distribution as the quantity



# Mathematical view

- Choose a random sample from population, the value of sample has the same distribution as  $X$
- Take  $n$  sample,  $i$ -th sample = Random variable  $X_i$  (has the same distribution as  $X$ )
- $X_1, X_2, \dots, X_n$ : independent and have same distribution as  $X$  = *independent identically distributed* (i.i.d.) RVs

# Describing data

- Using frequency table and graph
- Graphs: line, bar, polygon

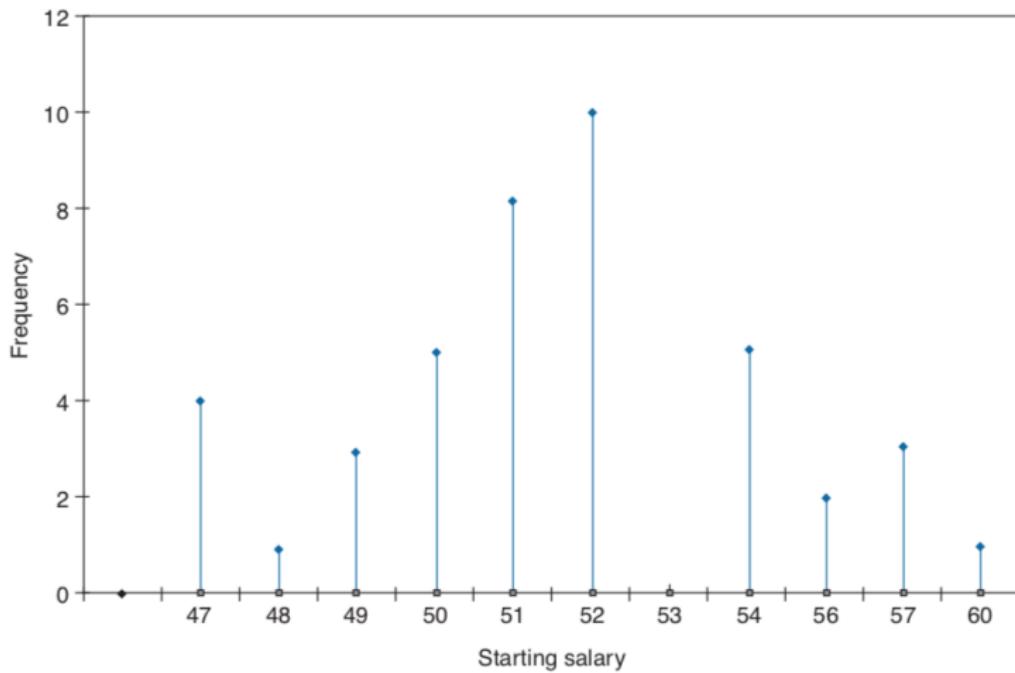
# Table

TABLE 2.1 *Starting Yearly Salaries*

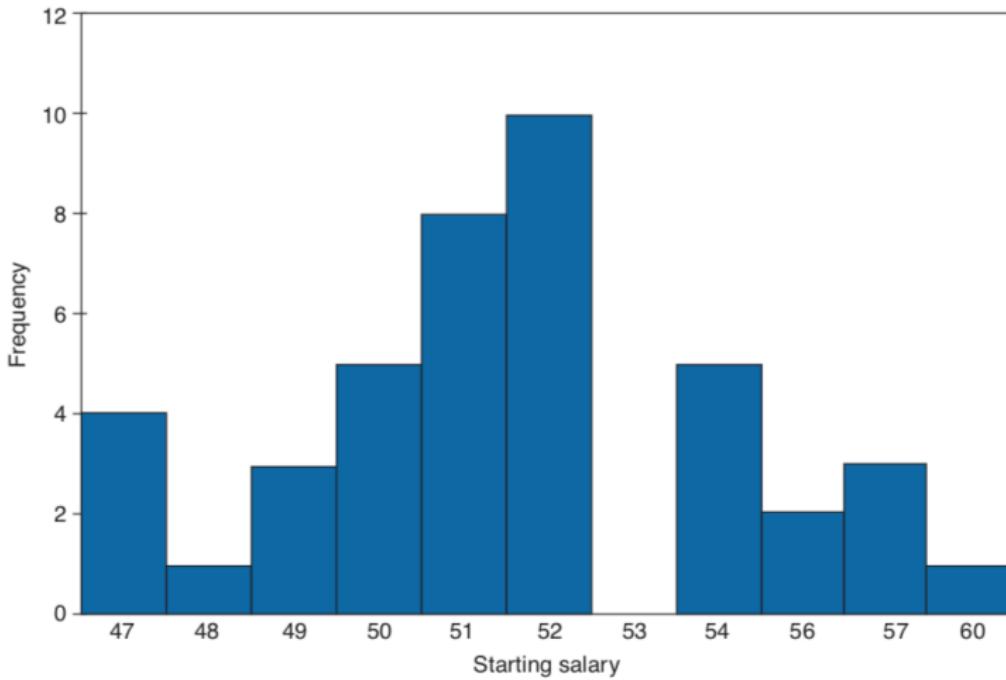
Starting Salary	Frequency
47	4
48	1
49	3
50	5
51	8
52	10
53	0
54	5
56	2
57	3
60	1



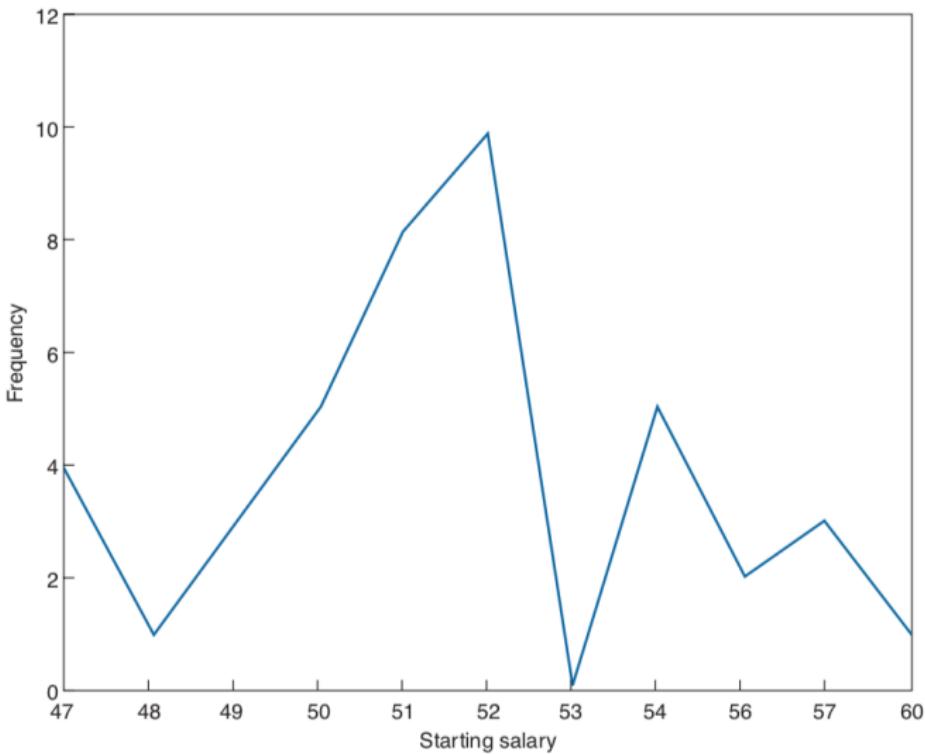
# Line graph



# Bar graph



# Polygon



# Pie chart

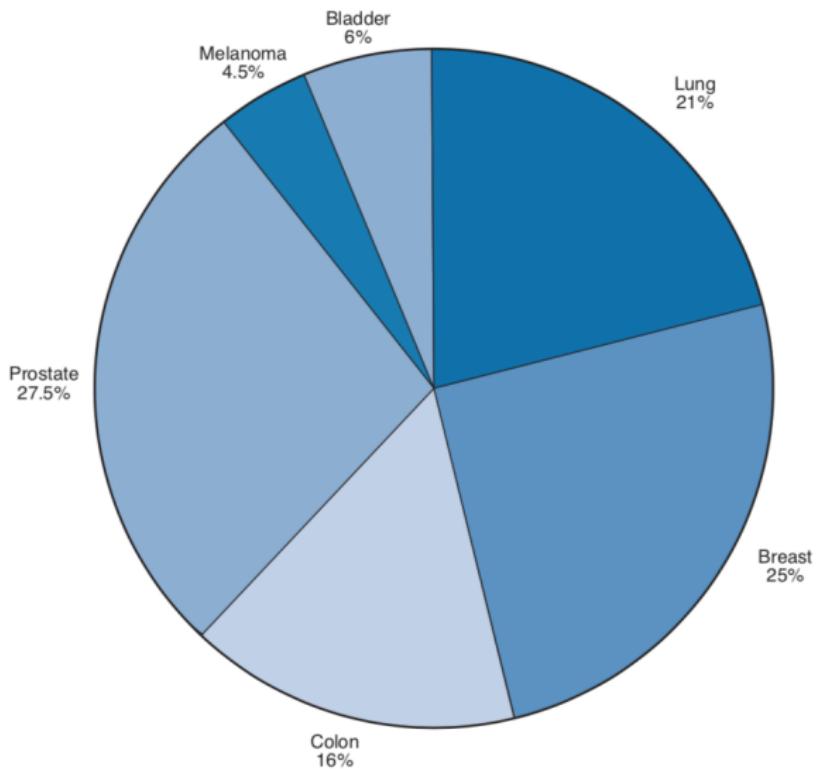
- Can use relative frequency (percentage)
- Presenting by pie chart



# Relative frequency table

Type of Cancer	Number of New Cases	Relative Frequency
Lung	42	.21
Breast	50	.25
Colon	32	.16
Prostate	55	.275
Melanoma	9	.045
Bladder	12	.06

# Pie chart



# Histogram

- For large data we don't use direct graph
- Divide data into intervals
- Count frequency of data in each interval
- Bar graph of intervals = histogram

# Raw data

TABLE 2.3 *Life in Hours of 200 Incandescent Lamps*

Item Lifetimes									
1,067	919	1,196	785	1,126	936	918	1,156	920	948
855	1,092	1,162	1,170	929	950	905	972	1,035	1,045
1,157	1,195	1,195	1,340	1,122	938	970	1,237	956	1,102
1,022	978	832	1,009	1,157	1,151	1,009	765	958	902
923	1,333	811	1,217	1,085	896	958	1,311	1,037	702
521	933	928	1,153	946	858	1,071	1,069	830	1,063
930	807	954	1,063	1,002	909	1,077	1,021	1,062	1,157
999	932	1,035	944	1,049	940	1,122	1,115	833	1,320
901	1,324	818	1,250	1,203	1,078	890	1,303	1,011	1,102
996	780	900	1,106	704	621	854	1,178	1,138	951
1,187	1,067	1,118	1,037	958	760	1,101	949	992	966
824	653	980	935	878	934	910	1,058	730	980
844	814	1,103	1,000	788	1,143	935	1,069	1,170	1,067
1,037	1,151	863	990	1,035	1,112	931	970	932	904
1,026	1,147	883	867	990	1,258	1,192	922	1,150	1,091
1,039	1,083	1,040	1,289	699	1,083	880	1,029	658	912
1,023	984	856	924	801	1,122	1,292	1,116	880	1,173
1,134	932	938	1,078	1,180	1,106	1,184	954	824	529
998	996	1,133	765	775	1,105	1,081	1,171	705	1,425
610	916	1,001	895	709	860	1,110	1,149	972	1,002

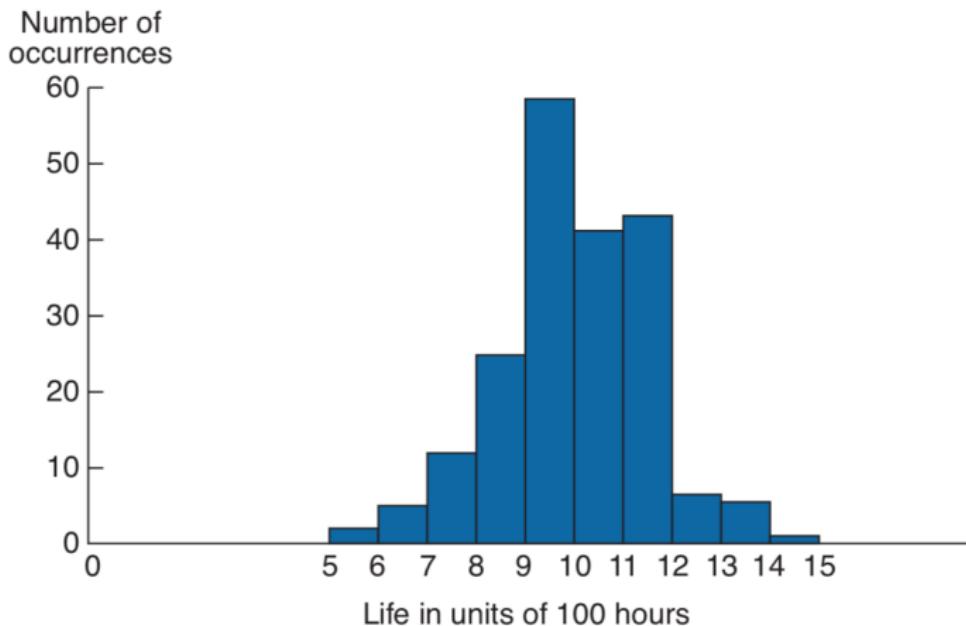


# Interval frequency

TABLE 2.4 *A Class Frequency Table*

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

# Histogram



# Important Statistics



# Sample mean

- Data:  $x_1, x_2, \dots, x_n$
- Sample mean:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- If  $y_i = ax_i + b$  for all  $i$  then

$$\bar{y} = a\bar{x} + b$$

# Weighted mean

- Data:  $x_i$  has frequency  $f_i$ ,  $i = 1, \dots, k$   
and  $\sum_{i=1}^k f_i = n$
- $\bar{x} = \frac{1}{n} \sum_{i=1}^k x_i f_i$

# Sample median

- Sorted data:  $x_1 \leq x_2 \leq \cdots \leq x_n$
- If  $n$  is odd, median is  $x_{(n+1)/2}$
- If  $n$  is even, median is  $\frac{1}{2}(x_{n/2} + x_{n/2+1})$
- Median is the middle number

# Sample mode

- Sorted data:  $x_1 \leq x_2 \leq \cdots \leq x_n$
- Mode = number has largest frequency
- Mode = most

# Distribution of Sample mean

# Population

- Suppose we want to investigate a quantity from a large population
- Ex: the height of men in a city
- Assume that there is a underlying random variable  $X$  that have the same distribution as the quantity

- Suppose  $X$  is a random variable with distribution  $F$ , mean  $\mu$  and variance  $\sigma^2$
- Think of  $X$  as the "real" distribution,  $\mu$  and  $\sigma^2$  as "real" mean and variance
- $E(X) = \mu$ : population mean,  
 $\text{Var}(X) = \sigma^2$ : population variance
- We want to find  $F$  and  $\mu$  and  $\sigma^2$

# Sample

- If we observe  $n$  samples  $X_1, X_2, \dots, X_n$  choose from the population in a “random way”
- then  $X_i$ 's are independent random variables with same distribution as  $X$
- We will use samples to understand the hidden distribution  $F$  of  $X$

# Statistics from sample

- A random variable  $T$  is called a *statistics* of the sample if  $T$  is a function of  $X_1, \dots, X_n$ .

$$T = f(X_1, X_2, \dots, X_n)$$

- $T$  is a random variable, we want to understand the distribution of  $T$
- 2 important statistics:  $T$  is Sample mean and  $T$  is sample variance



# Sample mean

- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$
- $\bar{X}$  is a function of the samples = a random variable
- want to find the distribution of  $\bar{X}$



# Mean

$$\begin{aligned}E(\bar{X}) &= \frac{1}{n} \sum_{i=1}^n E(X_i) \\&= \frac{1}{n} n\mu = \mu\end{aligned}$$

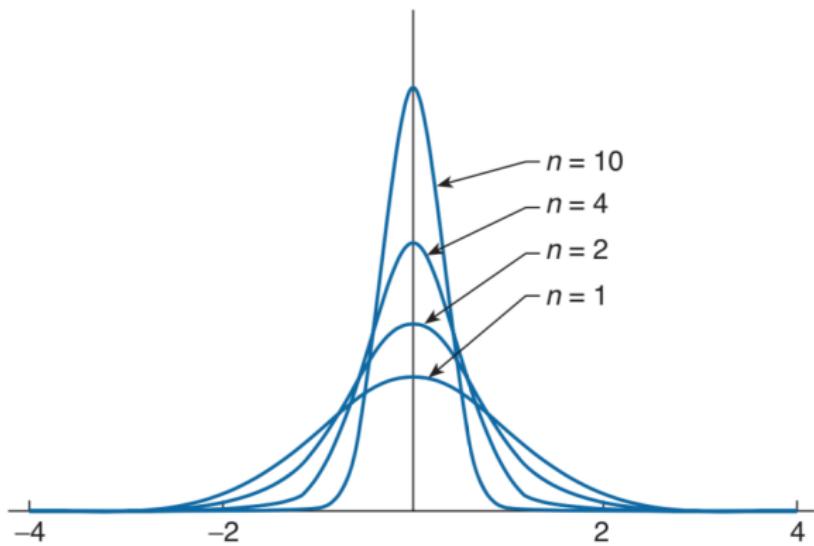
- $\bar{X}$  has the same mean as  $X$ .
- $\bar{X}$  is used to estimate  $\mu$  and called a *unbias estimator* of  $\mu$

# Variance

$$\begin{aligned}\text{Var}(\bar{X}) &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \\ &= \frac{\sigma^2}{n}\end{aligned}$$

- $\bar{X}$  spreads much smaller than  $X$ .
- When  $n$  large  $\bar{X}$  concentrates tightly around  $\mu$ .





# Central limit theorem

- Suppose  $X_1, \dots, X_n$  independent identically distributed RVs with mean  $\mu$  and variance  $\sigma^2$ .
- then  $X_1 + \dots + X_n$  has distribution approximately **normal** with mean  $n\mu$  and variance  $n\sigma^2$

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0, 1)$$

# Distribution of $\bar{X}$

- $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$
- $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$
- $\bar{X}$  has distribution approximately normal with mean  $\mu$  and variance  $\frac{\sigma^2}{n}$ .

# Example

The weights (in lbs) of a population of workers have mean 167 and standard deviation 27.

- (a) If a sample of 36 workers is chosen, approximate the probability that the sample mean of their weights lies between 163 and 170.
- (b) Repeat part (a) when the sample is of size 144.

# Solution

(a)  $\bar{X} \sim N(167, 27/\sqrt{36})$

$$\begin{aligned} P\{163 < \bar{X} < 170\} &= P\left\{\frac{163 - 167}{4.5} < \frac{\bar{X} - 167}{4.5} < \frac{170 - 167}{4.5}\right\} \\ &= P\left\{-.8889 < \frac{\bar{X} - 167}{4.5} < .8889\right\} \\ &\approx 2P\{Z < .8889\} - 1 \\ &\approx .6259 \end{aligned}$$



# Solution

(b)  $\bar{X} \sim N(167, 27/\sqrt{144})$

$$\begin{aligned} P\{163 < \bar{X} < 170\} &= P\left\{\frac{163 - 167}{2.25} < \frac{\bar{X} - 167}{2.25} < \frac{170 - 167}{2.25}\right\} \\ &= P\left\{-1.7778 < \frac{\bar{X} - 167}{4.5} < 1.7778\right\} \\ &\approx 2P\{Z < 1.7778\} - 1 \\ &\approx .9246 \end{aligned}$$

# Example

An astronomer wants to measure the distance to a distant star. She makes a series of measurements and then use their average value as an estimate of  $d$ . If the values of the measurements are independent random variables with mean of  $d$  light years and standard deviation of 2 light years, how many measurements need she make to be at least 95 percent certain that her estimate is accurate to within  $\pm 0.5$  light years?

# Solution

$$\bar{X} \sim N(d, 2/\sqrt{n})$$

$$\begin{aligned} P\{-.5 < \bar{X} - d < .5\} &= P\left\{\frac{-.5}{2/\sqrt{n}} < \frac{\bar{X} - d}{2/\sqrt{n}} < \frac{.5}{2/\sqrt{n}}\right\} \\ &\approx P\{-\sqrt{n}/4 < Z < \sqrt{n}/4\} \\ &= 2P\{Z < \sqrt{n}/4\} - 1 \end{aligned}$$



Need

$$2P(Z < \sqrt{n}/4) - 1 \geq 0.95$$

or

$$\Phi(\sqrt{n}/4) \geq 0.975$$

so  $\sqrt{n}/4 \geq 1.96$  or

$$n \geq 62$$



# Practice

Civil engineers believe that  $W$ , the amount of weight (in units of 1,000 pounds) that a certain bridge can withstand without being damaged, is normally distributed with mean 400 and standard deviation 40. Suppose that the weight (in units of 1,000 pounds) of a car is a random variable with mean 3 and standard deviation .3. How many cars would have to be on the bridge for the probability of damage to exceed .1?



# Sample variance

# Sample variance

- $X_1, \dots, X_n$  sample of  $X$
- Statistic  $S^2$  is called sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- $S = \sqrt{S^2}$ : sample standard deviation



# Mean

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 \\&= \sum_{i=1}^n X_i^2 - 2\bar{X} \sum_{i=1}^n X_i + n\bar{X}^2 \\&= \sum_{i=1}^n X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2\end{aligned}$$

We'll use the fact:  $E(X^2) = \text{Var}(X) + E(X)^2$

# Take expectation both sides

$$\begin{aligned}(n - 1)E[S^2] &= E \left[ \sum_{i=1}^n X_i^2 \right] - nE[\bar{X}^2] \\&= nE[X_1^2] - nE[\bar{X}^2] \\&= n\text{Var}(X_1) + n(E[X_1])^2 - n\text{Var}(\bar{X}) - n(E[\bar{X}])^2 \\&= n\sigma^2 + n\mu^2 - n(\sigma^2/n) - n\mu^2 \\&= (n - 1)\sigma^2\end{aligned}$$

so  $E(S^2) = \sigma^2$

$S^2$  is used to estimate  $\sigma^2$  and called a *unbiased estimator* of  $\sigma^2$



# Sampling from Normal population



# Sample mean

- $X \sim N(\mu, \sigma^2)$
- $X_1, \dots, X_n$  sample of  $X$
- $(X_1 + \dots + X_n) \sim N(n\mu, n\sigma^2)$
- $\bar{X} \sim N(\mu, \sigma^2/n)$

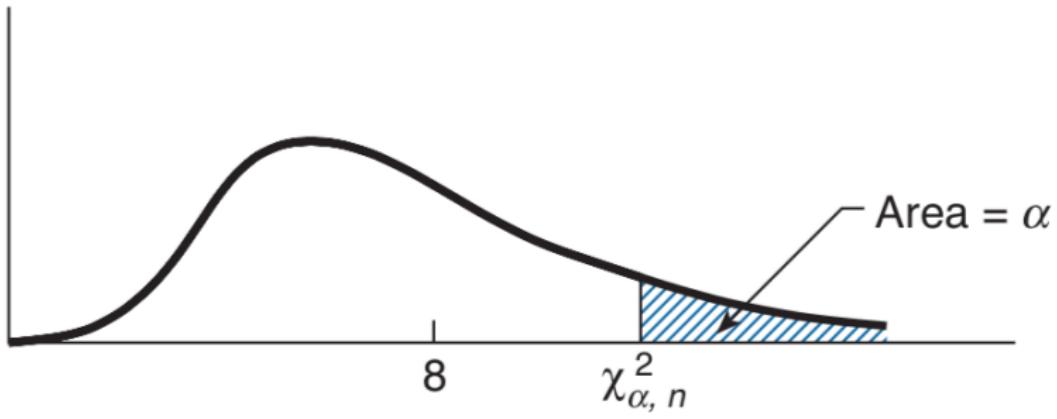


# Chi-square RV

- $X_1, \dots, X_n$  i.i.d.  $\mathbf{N}(0, 1)$
- $Y = X_1^2 + \dots + X_n^2$  is said to have the chi-square distribution with  $n$  degree of freedom
- Denote  $Y \sim \chi_n^2$
- Define  $\chi_{\alpha,n}^2$  to be a number so that

$$P(Y > \chi_{\alpha,n}^2) > \alpha$$





# Sum of chi square

- If  $X \sim \chi_n^2$  and  $Y \sim \chi_m^2$
- and  $X, Y$  independent
- then  $X + Y \sim \chi_{n+m}^2$



# Sample Variance

- $\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2$
- $$\sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} + \left( \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \right)^2$$



$$\sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2$$

- $\sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2$
- $\left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2 \sim \chi_1^2$
- $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$  assuming  $\bar{X}$  and  $S^2$  independent



# Samples from finite population



# Ratio in finite population

- Population of a quantity in real world is never truly infinity
- Size of population = a finite number  $N$
- Let  $p$  be a number in  $(0, 1)$
- Suppose in a population of size  $N$  there are  $pN$  elements that have a certain property A
- We want to find the value of  $p$

# Dependent Samples

- Take  $n$  samples from population of size  $N$
- $X_i = 1$  if the  $i$ -th sample has property A
- $X_i = 0$  otherwise
- then  $P(X_i = 1) = p$  for all  $i$
- but  $X_1, \dots, X_n$  are not independent

# Almost independent

- $X = X_1 + \cdots + X_n$ : total number of samples have property A
- If  $N$  is large,  $n$  is small then  $X_1, \dots, X_n$  are almost independent
- then  $X$  is almost binomial  $(n, p)$
- then  $X$  is almost normal  
 $\mathcal{N}(np, np(1 - p))$

# Estimate sample mean

- $\bar{X} = \frac{X}{n}$
- $\bar{X}$  is almost normal  $\mathcal{N}(p, p(1 - p)/n)$
- Can use  $p$  and  $n$  to calculate probability about  $\bar{X}$  and  $X$



# Example

Suppose that 45 percent of the population favors a certain candidate in an upcoming election. If a random sample of size 200 is chosen, find

- (a) the expected value and standard deviation of the number of members of the sample that favor the candidate;
- (b) the probability that more than half the members of the sample favor the candidate.



# Solution

- $E(X) = np = 200(.45) = 90$
- $SD(X) = \sqrt{200(.45)(.55)} = 7.0356$

$$\begin{aligned}P(X > 100) &= P\left(\frac{X - 90}{7.0356} \geq \frac{100.5 - 90}{7.0356}\right) \\&\approx P(Z > 1.4924) = 0.0678\end{aligned}$$

# Homework

Chapter 6: 2, 3, 6, 8, 12, 18, 19



# Practice

Fifty-two percent of the residents of a certain city are in favor of teaching evolution in high school. Find or approximate the probability that at least 50 percent of a random sample of size  $n$  is in favor of teaching evolution, when

- (a)  $n = 10$ ;
- (b)  $n = 100$ ;
- (c)  $n = 1,000$ ;
- (d)  $n = 10,000$ .

# Practice

An instructor knows that student exam scores have mean 77 and standard deviation 15. He has two classes of size 25 and 64.

- (a) Approximate the probability that the average test score in the class of size 25 lies between 72 and 82.
- (b) Repeat part (a) for a class of size 64.
- (c) Approximate probability that the average test score in the class of size 25 is higher than that of the class of size 64?