



## Lecture 3

**Sampling distributions.** Counts, Proportions, and sample mean.

- **Statistical Inference:** Uses data and summary statistics (mean, variances, proportions, slopes) to draw ***conclusions*** about a population or process.
- **Statistic:** Any random variable measured from a random sample or in a random experiment.
- **Sampling distribution of a statistic:** shows how a statistic varies in repeated measurements of an experiment. The probability distribution of a statistic is called its sampling distribution.
- **Population distribution of a statistic:** distribution of values for all members of the population. Unknown, but estimable using laws of statistics.

# ***Sampling Distribution for Counts and Proportions:***

- In a survey of 2500 engineers, 600 of them say they would consider working as a consultant. Let  $X$  = the number who would work as consultants.
- $X$  is a count:
- Sample Proportion of people who would work as consultants:

Distinguish count from sample proportion, they have *different* distributions.

# Binomial Distribution for Sample Counts:

- Distribution of the count,  $X$ , of successes in a binomial setting with parameters  $n$  and  $p$
- $n$  = number of observations
- $p$  =  $P$  (Success) on any one observation
- $X$  can take values from 0 to  $n$
- **Notation:**  $X \sim \text{Bin} (n, p)$
- **Setting:**
  1. Fixed number of  $n$  observations
  2. All observations are independent of each other
  3. Each observation falls into one of two categories: Success or Failure
  4.  $P$  (Success) =  $P$  (S) =  $p$

# EXAMPLES (Bin or not Bin)

- Toss a fair coin 10 times and count the number  $X$  of heads. What about a biased coin?
- Deal 10 cards from a shuffled deck of 52.  $X$  is the number of spades. Suggestions??
- Number of girls born among first 100 children in a (large) hospital this year.
- Number of girls born in this hospital so far this year.

# Finding Binomial Probabilities

Use Table C: page T-6

- (How to: - find your  $n$  = number of observations
- find your  $p$  = probability of success
- find the probability corresponding to  $k$  = number of successes you are interested in)
- You can use R as well to evaluate probabilities:
  - » `pbinom(4,size=10,prob=0.15)` (calculates  $P(\text{Bin}(10,0.15) \leq 4)$  )
  - » [1] 0.990126
- If you want the entry in the table do:
  - » `pbinom(4,size=10,prob=0.15)-pbinom(3,size=10,prob=0.15)`
  - » [1] 0.04009571



TABLE C Binomial probabilities (continued)

		Entry is $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$								
		<i>p</i>								
<i>n</i>	<i>k</i>	.10	.15	.20	.25	.30	.35	.40	.45	.50
2	0	.8100	.7225	.6400	.5625	.4900	.4225	.3600	.3025	.2500
	1	.1800	.2550	.3200	.3750	.4200	.4550	.4800	.4950	.5000
	2	.0100	.0225	.0400	.0625	.0900	.1225	.1600	.2025	.2500
3	0	.7290	.6141	.5120	.4219	.3430	.2746	.2160	.1664	.1250
	1	.2430	.3251	.3840	.4219	.4410	.4436	.4320	.4084	.3750
	2	.0270	.0574	.0960	.1406	.1890	.2389	.2880	.3341	.3750
	3	.0010	.0034	.0080	.0156	.0270	.0429	.0640	.0911	.1250
4	0	.6561	.5220	.4096	.3164	.2401	.1785	.1296	.0915	.0625
	1	.2916	.3685	.4096	.4219	.4116	.3845	.3456	.2995	.2500
	2	.0486	.0975	.1536	.2109	.2646	.3105	.3456	.3675	.3750
	3	.0036	.0115	.0256	.0469	.0756	.1115	.1536	.2005	.2500
	4	.0001	.0005	.0016	.0039	.0081	.0150	.0256	.0410	.0625
5	0	.5905	.4437	.3277	.2373	.1681	.1160	.0778	.0503	.0313
	1	.3280	.3915	.4096	.3955	.3602	.3124	.2592	.2059	.1563
	2	.0729	.1382	.2048	.2637	.3087	.3364	.3456	.3369	.3125
	3	.0081	.0244	.0512	.0879	.1323	.1811	.2304	.2757	.3125
	4	.0004	.0022	.0064	.0146	.0284	.0488	.0768	.1128	.1562
	5		.0001	.0003	.0010	.0024	.0053	.0102	.0185	.0312
6	0	.5314	.3771	.2621	.1780	.1176	.0754	.0467	.0277	.0156
	1	.3543	.3993	.3932	.3560	.3025	.2437	.1866	.1359	.0938
	2	.0984	.1762	.2458	.2966	.3241	.3280	.3110	.2780	.2344
	3	.0146	.0415	.0819	.1318	.1852	.2355	.2765	.3032	.3125
	4	.0012	.0055	.0154	.0330	.0595	.0951	.1382	.1861	.2344
	5	.0001	.0004	.0015	.0044	.0102	.0205	.0369	.0609	.0937
	6			.0001	.0002	.0007	.0018	.0041	.0083	.0156
7	0	.4783	.3206	.2097	.1335	.0824	.0490	.0280	.0152	.0078
	1	.3720	.3960	.3670	.3115	.2471	.1848	.1306	.0872	.0547
	2	.1240	.2097	.2753	.3115	.3177	.2985	.2613	.2140	.1641
	3	.0230	.0617	.1147	.1730	.2269	.2679	.2903	.2918	.2734
	4	.0026	.0109	.0287	.0577	.0972	.1442	.1935	.2388	.2734
	5	.0002	.0012	.0043	.0115	.0250	.0466	.0774	.1172	.1641
	6		.0001	.0004	.0013	.0036	.0084	.0172	.0320	.0547
	7				.0001	.0002	.0006	.0016	.0037	.0078
8	0	.4305	.2725	.1678	.1001	.0576	.0319	.0168	.0084	.0039
	1	.3826	.3847	.3355	.2670	.1977	.1373	.0896	.0548	.0313
	2	.1488	.2376	.2936	.3115	.2965	.2587	.2090	.1569	.1094
	3	.0331	.0839	.1468	.2076	.2541	.2786	.2787	.2568	.2188
	4	.0046	.0185	.0459	.0865	.1361	.1875	.2322	.2627	.2734
	5	.0004	.0026	.0092	.0231	.0467	.0808	.1239	.1719	.2188
	6		.0002	.0011	.0038	.0100	.0217	.0413	.0703	.1094
	7			.0001	.0004	.0012	.0033	.0079	.0164	.0312
	8					.0001	.0002	.0007	.0017	.0039





TABLE C Binomial probabilities (continued)

<i>n</i>	<i>k</i>	<i>p</i>								
		.10	.15	.20	.25	.30	.35	.40	.45	.50
9	0	.3874	.2316	.1342	.0751	.0404	.0207	.0101	.0046	.0020
	1	.3874	.3679	.3020	.2253	.1556	.1004	.0605	.0339	.0176
	2	.1722	.2597	.3020	.3003	.2668	.2162	.1612	.1110	.0703
	3	.0446	.1069	.1762	.2336	.2668	.2716	.2508	.2119	.1641
	4	.0074	.0283	.0661	.1168	.1715	.2194	.2508	.2600	.2461
	5	.0008	.0050	.0165	.0389	.0735	.1181	.1672	.2128	.2461
	6	.0001	.0006	.0028	.0087	.0210	.0424	.0743	.1160	.1641
	7			.0003	.0012	.0039	.0098	.0212	.0407	.0703
	8				.0001	.0004	.0013	.0035	.0083	.0176
	9						.0001	.0003	.0008	.0020
10	0	.3487	.1969	.1074	.0563	.0282	.0135	.0060	.0025	.0010
	1	.3874	.3474	.2684	.1877	.1211	.0725	.0403	.0207	.0098
	2	.1937	.2759	.3020	.2816	.2335	.1757	.1209	.0763	.0439
	3	.0574	.1298	.2013	.2503	.2668	.2522	.2150	.1665	.1172
	4	.0112	.0401	.0881	.1460	.2001	.2377	.2508	.2384	.2051
	5	.0015	.0085	.0264	.0584	.1029	.1536	.2007	.2340	.2461
	6	.0001	.0012	.0055	.0162	.0368	.0689	.1115	.1596	.2051
	7		.0001	.0008	.0031	.0090	.0212	.0425	.0746	.1172
	8			.0001	.0004	.0014	.0043	.0106	.0229	.0439
	9					.0001	.0005	.0016	.0042	.0098
	10							.0001	.0003	.0010
12	0	.2824	.1422	.0687	.0317	.0138	.0057	.0022	.0008	.0002
	1	.3766	.3012	.2062	.1267	.0712	.0368	.0174	.0075	.0029
	2	.2301	.2924	.2835	.2323	.1678	.1088	.0639	.0339	.0161
	3	.0852	.1720	.2362	.2581	.2397	.1954	.1419	.0923	.0537
	4	.0213	.0683	.1329	.1936	.2311	.2367	.2128	.1700	.1208
	5	.0038	.0193	.0532	.1032	.1585	.2039	.2270	.2225	.1934
	6	.0005	.0040	.0155	.0401	.0792	.1281	.1766	.2124	.2256
	7		.0006	.0033	.0115	.0291	.0591	.1009	.1489	.1934
	8		.0001	.0005	.0024	.0078	.0199	.0420	.0762	.1208
	9			.0001	.0004	.0015	.0048	.0125	.0277	.0537
	10					.0002	.0008	.0025	.0068	.0161
	11						.0001	.0003	.0010	.0029
	12							.0001	.0002	.0002
15	0	.2059	.0874	.0352	.0134	.0047	.0016	.0005	.0001	.0000
	1	.3432	.2312	.1319	.0668	.0305	.0126	.0047	.0016	.0005
	2	.2669	.2856	.2309	.1559	.0916	.0476	.0219	.0090	.0032
	3	.1285	.2184	.2501	.2252	.1700	.1110	.0634	.0318	.0139
	4	.0428	.1156	.1876	.2252	.2186	.1792	.1268	.0780	.0417
	5	.0105	.0449	.1032	.1651	.2061	.2123	.1859	.1404	.0916
	6	.0019	.0132	.0430	.0917	.1472	.1906	.2066	.1914	.1527
	7	.0003	.0030	.0138	.0393	.0811	.1319	.1771	.2013	.1964
	8		.0005	.0035	.0131	.0348	.0710	.1181	.1647	.1964
	9		.0001	.0007	.0034	.0116	.0298	.0612	.1048	.1527
	10			.0001	.0007	.0030	.0096	.0245	.0515	.0916
	11				.0001	.0006	.0024	.0074	.0191	.0417
	12					.0001	.0004	.0016	.0052	.0139
	13						.0001	.0003	.0010	.0032
	14							.0001	.0002	.0005
	15								.0001	.0001

# Example

Your job is to examine light bulbs on an assembly line. You are interested in finding the probability of getting a defective light bulb, after examining 10 light bulbs.

- Let  $X$  = number of defective light bulbs
  - $P(\text{defective}) = .15$
  - $N = 10$
1. Is this a binomial set up?
  2. What is the probability that you get at most 2 defective light bulbs?
  3. What is the probability that the number of defective light bulbs you find is greater than eight?
  4. What is the probability that you find between 3 and 5 defective light bulbs?

# Binomial Mean and Standard Deviation

$$\mu_x = np$$

$$\sigma_x^2 = np(1 - p)$$

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

Example: Find the mean and standard deviation of the previous problems

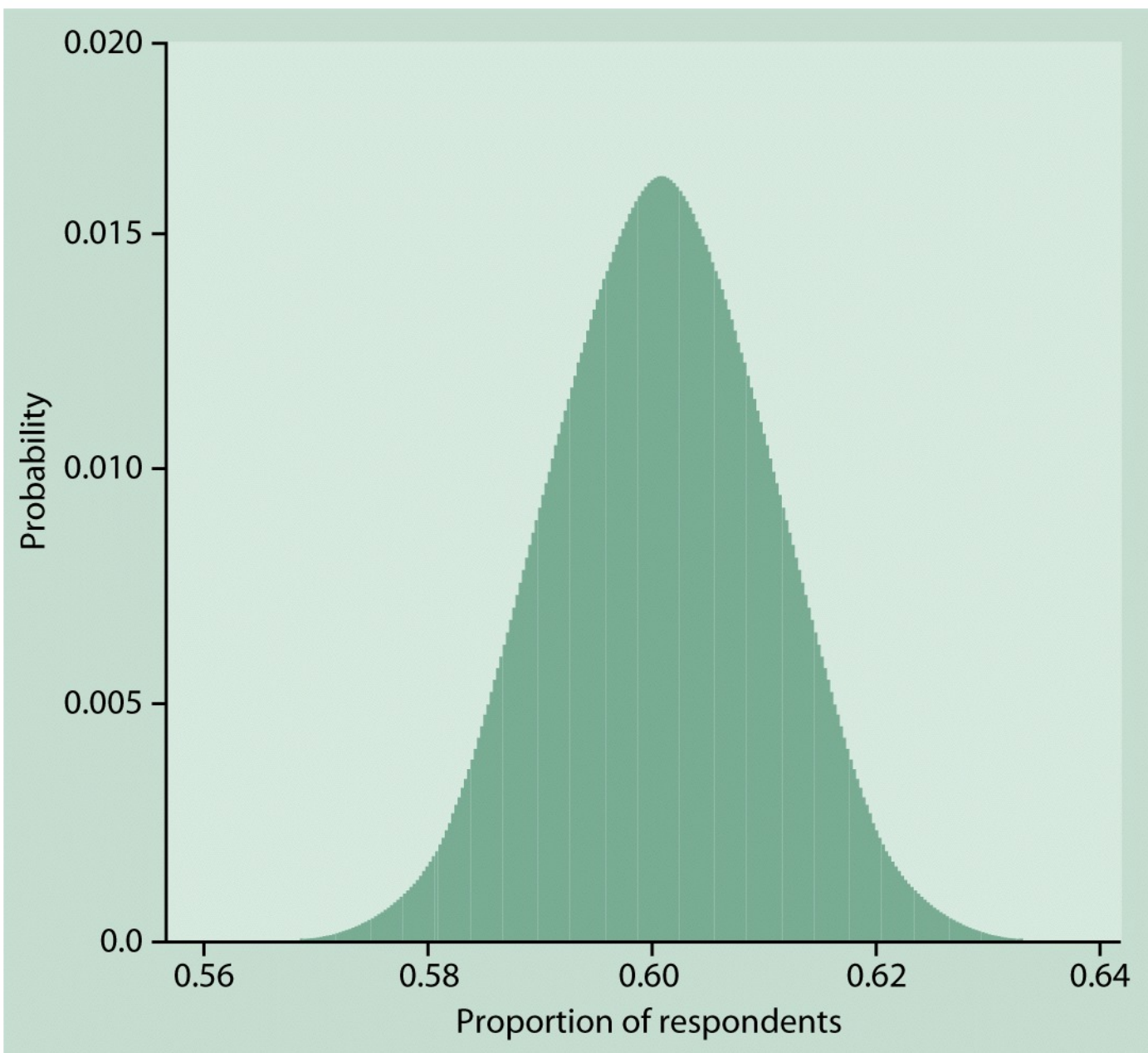
# Sample Proportions

- Let  $X$  be a count of successes in  $n =$  total number of observations in the data set.
- Then the sample proportion:

$$\hat{p} = \frac{X}{n}$$

– NOTE!!!!

- We know that  $X$  is distributed as a Binomial, however  $\hat{p}$  is NOT distributed as a Binomial.



## Normal approximation for counts and proportions

- If  $X$  is  $B(n,p)$ ,  $np \geq 10$  and  $n(1-p) \geq 10$  then:

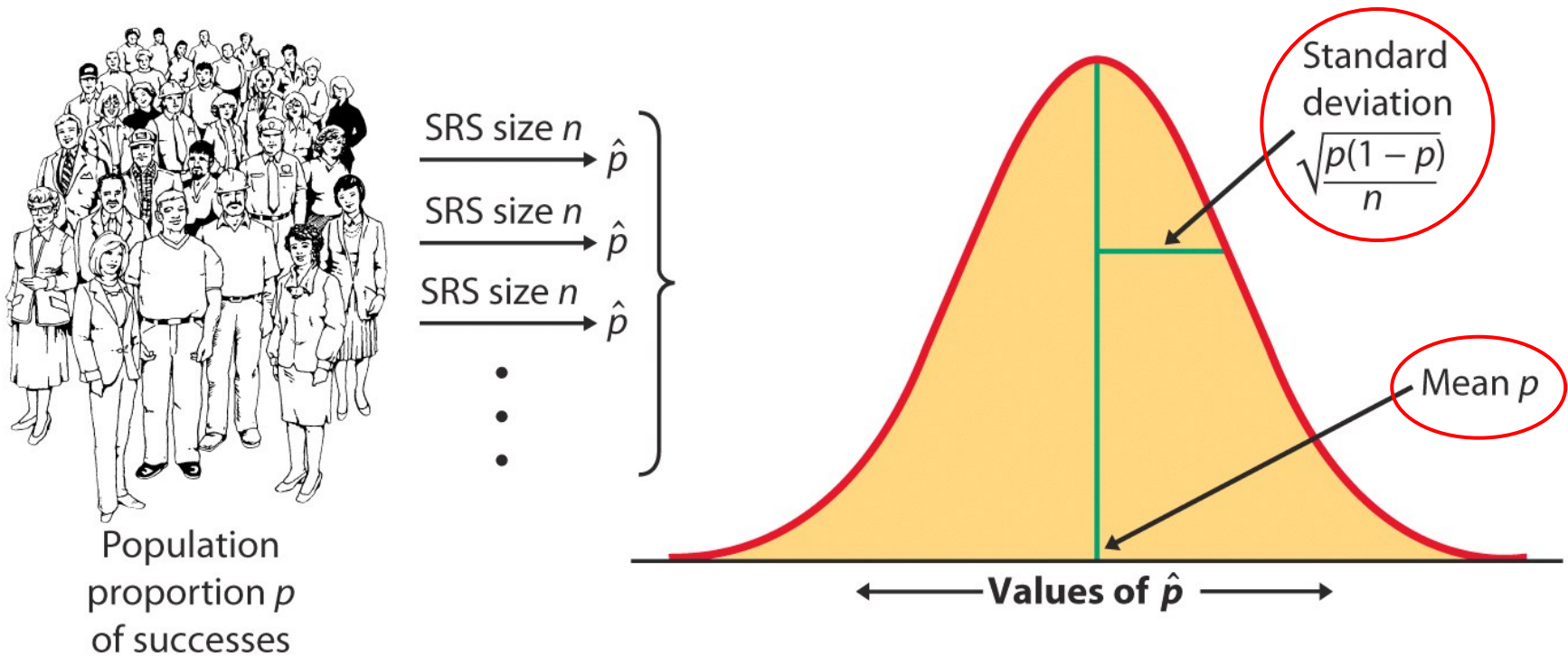
$X$  is approximately  $N( np, \sqrt{np(1-p)} )$

$\hat{p}$  is approximately  $N( p, \sqrt{\frac{p(1-p)}{n}} )$

# Sampling distribution of $\hat{p}$

The sampling distribution of  $\hat{p}$  is never exactly normal. But as the sample size increases, the sampling distribution of  $\hat{p}$  becomes approximately normal.

The normal approximation is most accurate for any fixed  $n$  when  $p$  is close to 0.5, and least accurate when  $p$  is near 0 or near 1.





# Example:

- In a survey 2500 engineers are asked if they would consider working as consultants. Suppose that 60% of the engineers would work as consultants. When we actually do the experiment 1375 say they would work as consultants

Find the mean and standard deviation of  $\hat{p}$ .

What is the probability that the percent of to be consultants in the sample is less than .58?

Between .59 and .61?

# The continuity correction:

- **Example:** According to a market research firm 52% of all residential telephone numbers in Los Angeles are unlisted. A telemarketing company uses random digit dialing equipment that dials residential numbers at random regardless of whether they are listed or not. The firm calls 500 numbers in L.A.
  1. What is the exact distribution of the number  $X$  of unlisted numbers that are called?
  2. Use a suitable approximation to calculate the probability that at least half the numbers are unlisted.

# The continuity correction(cont.):

- In the previous problem if we compute the probability that exactly 250 people had unlisted numbers using the normal approximation we would have find this probability equals zero.
- That is obviously not right because this number has to have some probability (small but still not zero).
- The problem comes from the fact that we use a continuous distribution (Normal Distribution) to approximate a discrete one (Binomial Distribution).
- So to improve the approximation we use a correction:
  - Whenever we compute a probability involving a count we will move the interval we compute 0.5 as to include or exclude the endpoints of the interval depending on the type of interval (closed or open) we compute in the problem.
  - Then we use the normal approximation to compute the probability of this new interval.

- Example: In the previous problem find:

$$P(X \geq 250)$$

$$P(X > 250)$$

$$P(X = 250)$$

$$P(X < 250)$$

$$P(248 < X < 251)$$

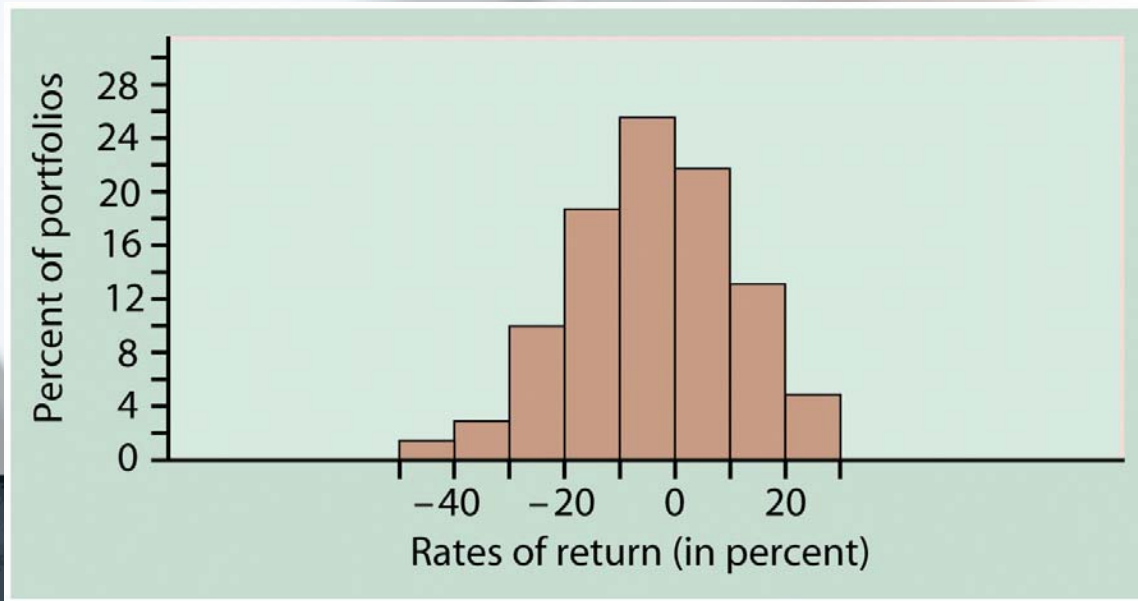
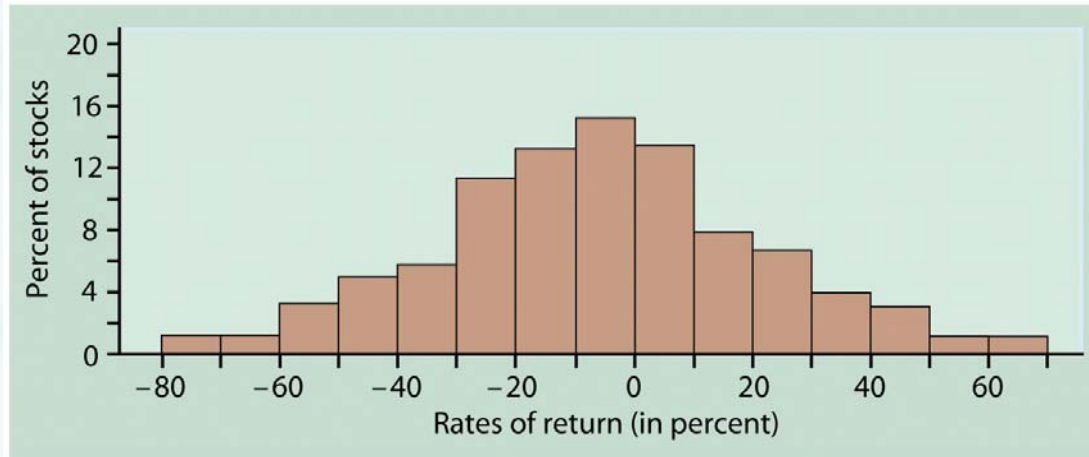
$$P(248 < X \leq 251)$$

$$P(248 \leq X \leq 251)$$

## ***Section 5.2: Sampling distribution of the sample mean***

- **Distribution of the center and spread**
- Setup:
- Draw a SRS (simple random sample) of size  $n$  from a population.
- Measure some variable  $X$  (i.e. income)
- Data:  $n$  random variables,  $X_1, X_2, X_3 \dots X_n$ , where  $X_i$  is a measurement on 1 individual (i.e. income of 1 individual in the sample)
- Since the individuals are randomly chosen, the  $X_i$ 's can be considered to be independent

# Example: Distribution of individual stocks (up) vs. distribution of mutual funds (down)

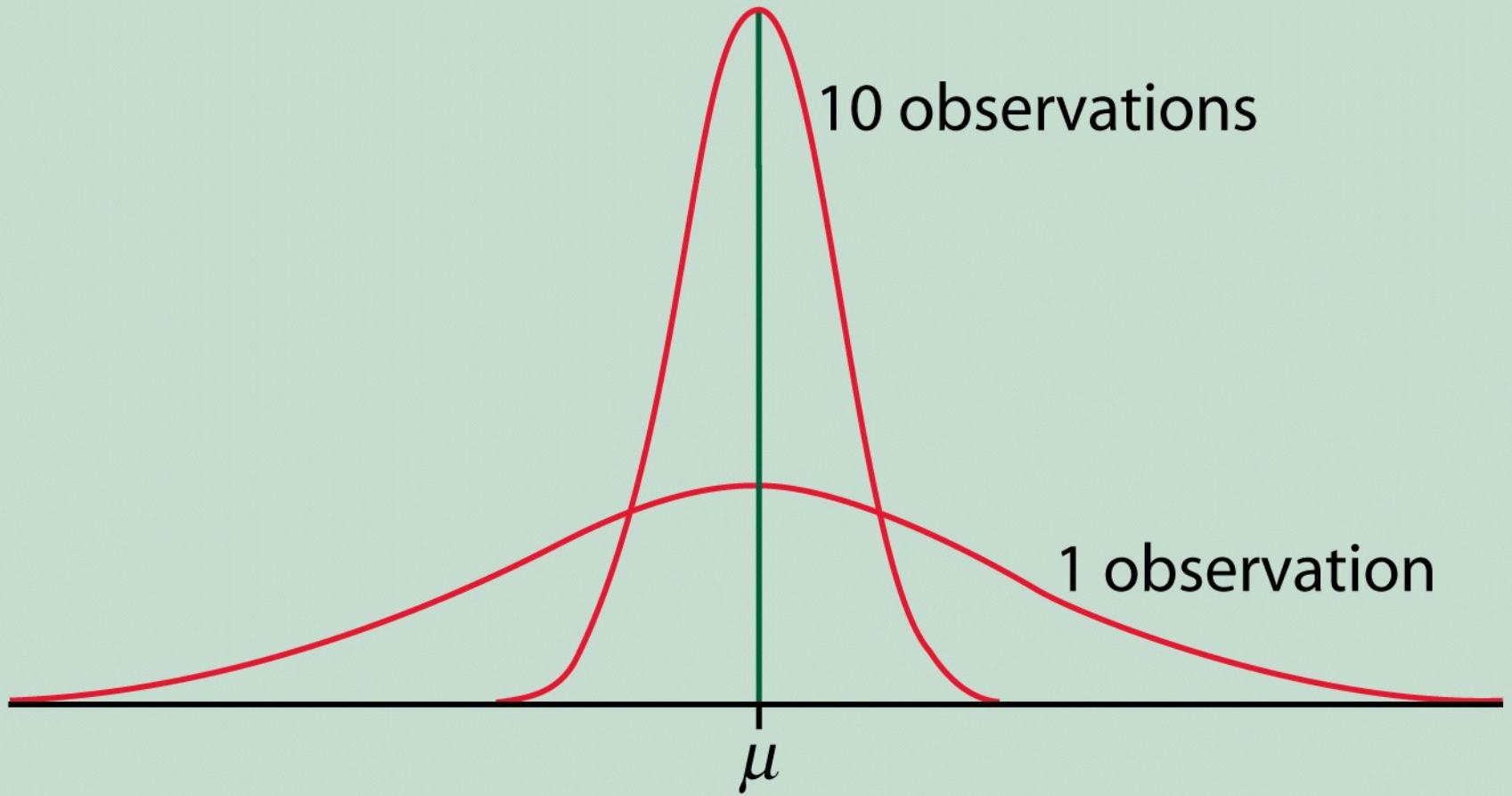


Sample mean:  $\bar{X} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n}$

- Let  $\bar{X}$  be the mean of an SRS (simple random sample) of size  $n$  from a population with mean  $\mu$  and standard deviation  $\sigma$ . The mean and standard deviation of  $\bar{X}$  are:

$$\mu_{\bar{X}} = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$





# Central Limit Theorem:

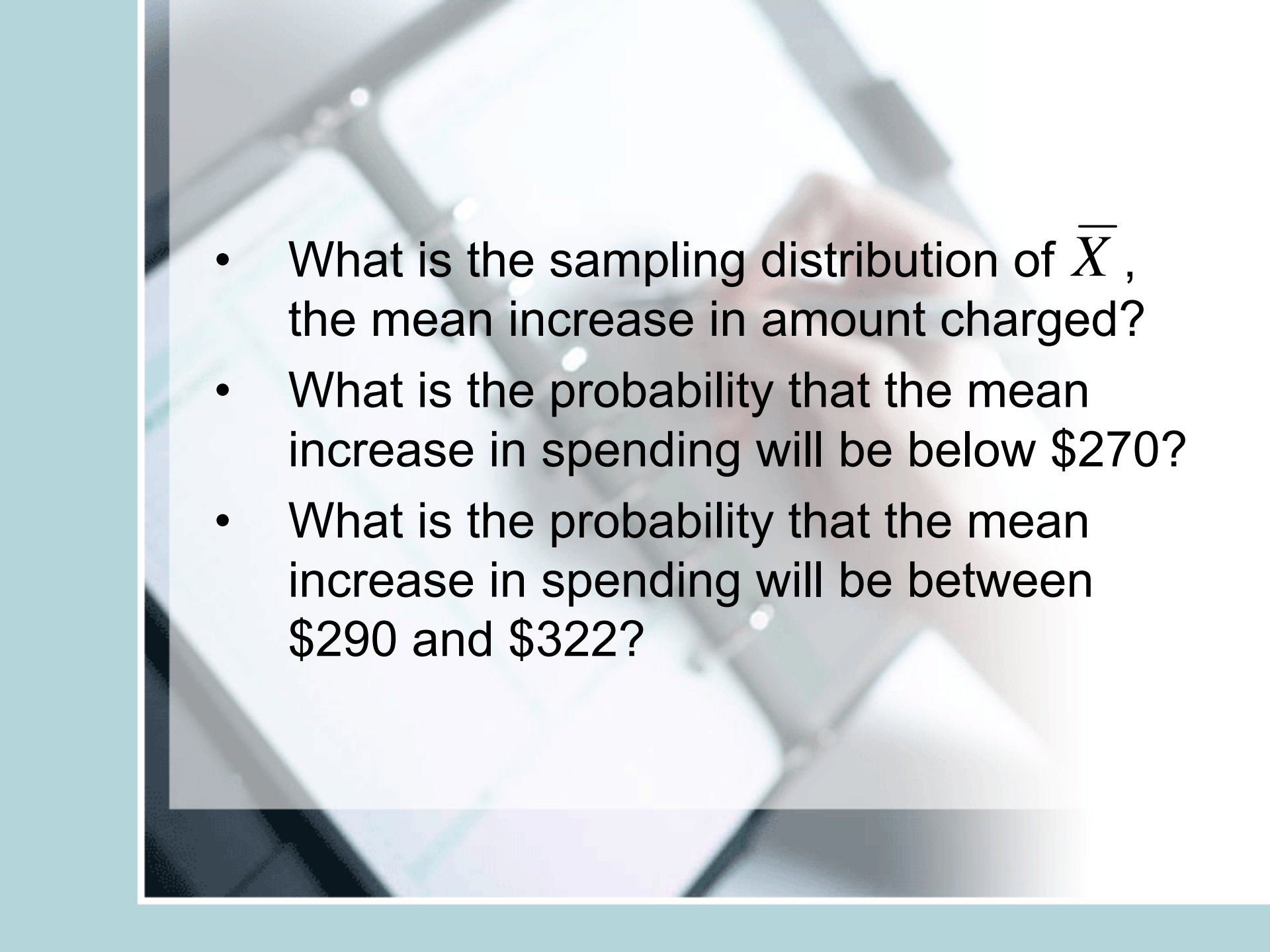
- Draw a SRS of size  $n$  ( $n$  large) from any population with mean  $\mu$  and standard deviation  $\sigma$ . The sampling distribution of the sample mean is approximately normal:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- **Important special case:** If the *population is normal* then the sample mean has exactly the normal distribution:  $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

## Example:

- A bank conducts an experiment to determine whether dropping their annual credit card fee will increase the amount charged on the credit card. The offer is made to a SRS of 200 customers. The bank then compares the amount the customers charged on their cards this year, to the amount charged next year. A mean increase of \$308 with a standard deviation of \$108 was found.

- 
- A hand holding a pen over a document with a grid pattern, likely a ledger or account book. The background is a light blue and white grid.
- What is the sampling distribution of  $\bar{X}$ , the mean increase in amount charged?
  - What is the probability that the mean increase in spending will be below \$270?
  - What is the probability that the mean increase in spending will be between \$290 and \$322?

## Example: 5.34

- The number of accidents per week at a hazardous intersection varies with mean 2.2 and standard deviation 1.4.
- What is the distribution of  $\bar{X}$ , the mean number of accidents in one year, (52 weeks)?
- What is the probability that  $\bar{X}$  is less than 2?
- What is the probability that there are fewer than 100 accidents in a year?

## Example: 5.67

- The weight of eggs produced by a certain breed of hen is Normally distributed with mean 65 grams and standard deviation 5 grams. Let cartons of such eggs be considered to be SRSs of size 12. What is the probability that the weight of a carton falls between 750 grams and 825 grams?

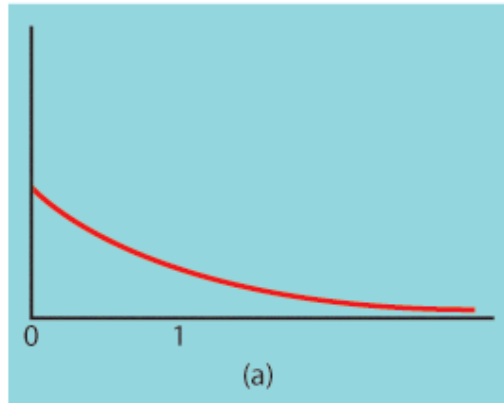
# Practical note

- ❑ Large samples are not always attainable.
  - ❑ Sometimes the cost, difficulty, or preciousness of what is studied drastically limits any possible sample size.
  - ❑ Blood samples/biopsies: No more than a handful of repetitions acceptable. Often, we even make do with just one.
  - ❑ Opinion polls have a limited sample size due to time and cost of operation. During election times, though, sample sizes are increased for better accuracy.
- ❑ Not all variables are normally distributed.
  - ❑ Income, for example, is typically strongly skewed.
  - ❑ Is  $\bar{x}$  still a good estimator of  $\mu$  then?

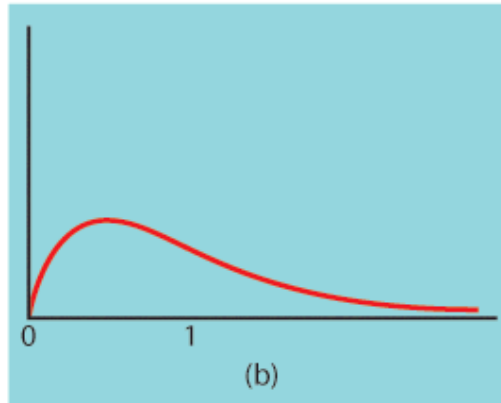
# The central limit theorem

**Central Limit Theorem:** When randomly sampling from any population with mean  $\mu$  and standard deviation  $\sigma$ , **when  $n$  is large enough**, the sampling distribution of  $\bar{x}$  is approximately normal:  $\sim N(\mu, \sigma/\sqrt{n})$ .

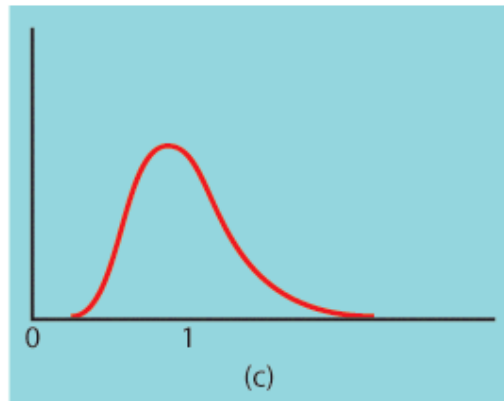
Population with strongly skewed distribution



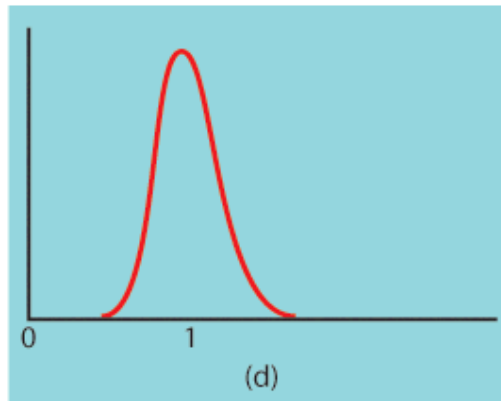
Sampling distribution of  $\bar{x}$  for  $n = 2$  observations



Sampling distribution of  $\bar{x}$  for  $n = 10$  observations



Sampling distribution of  $\bar{x}$  for  $n = 25$  observations

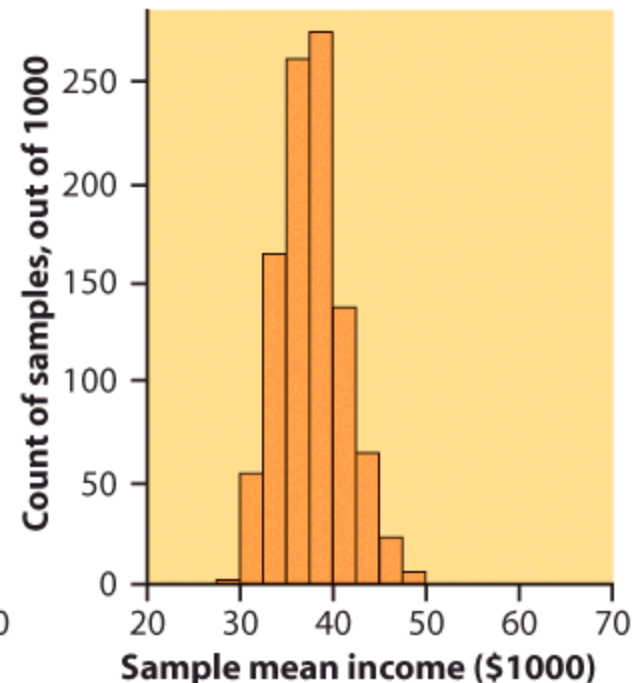
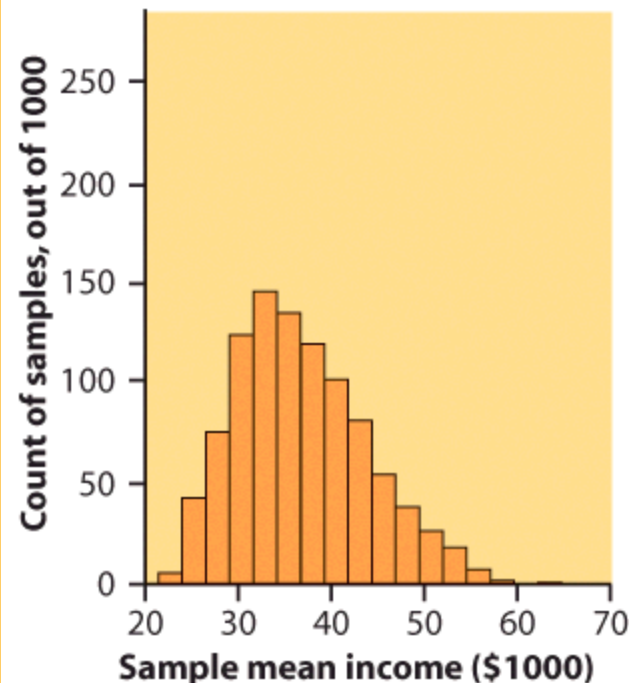


## Income distribution

Let's consider the very large database of individual incomes from the Bureau of Labor Statistics as our population. It is strongly right skewed.

- We take 1000 SRSs of 100 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.
- We also take 1000 SRSs of 25 incomes, calculate the sample mean for each, and make a histogram of these 1000 means.

Which histogram corresponds to the samples of size 100? 25?





# How large a sample size?

It depends on the population distribution. More observations are required if the population distribution is far from normal.

- ▣ A sample size of 25 is generally enough to obtain a normal sampling distribution from a strong skewness or even mild outliers.
- ▣ A sample size of 40 will typically be good enough to overcome extreme skewness and outliers.

*In many cases,  $n = 25$  isn't a huge sample. Thus, even for strange population distributions we can assume a normal sampling distribution of the mean and work with it to solve problems.*

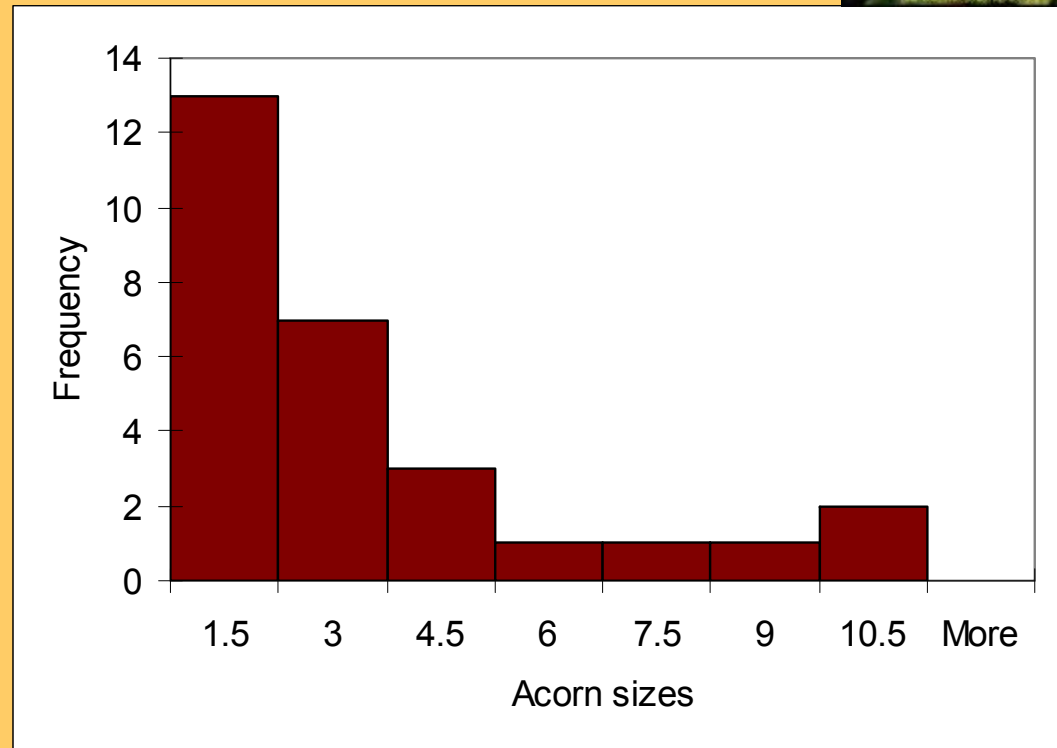
# Sampling distributions



Atlantic acorn sizes (in  $\text{cm}^3$ )

— sample of 28 acorns:

- Describe the histogram.  
What do you assume for the population distribution?



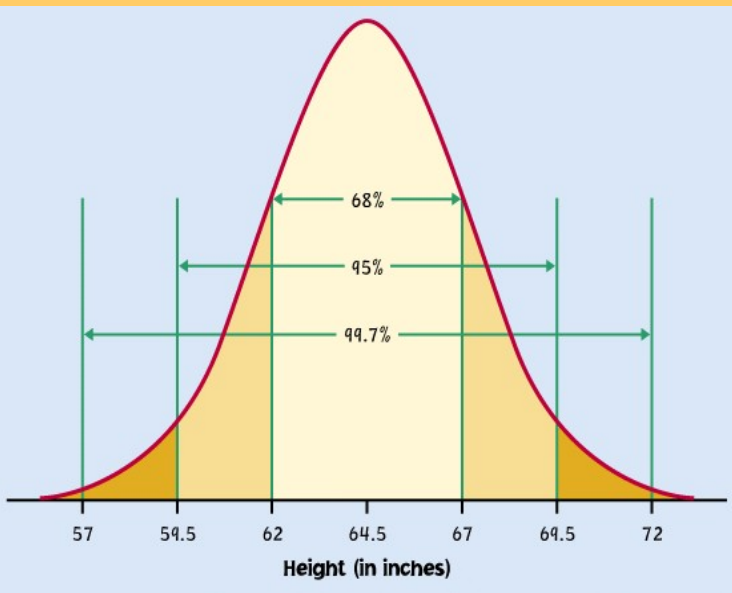
- What would be the shape of the sampling distribution of the mean:
  - For samples of size 5?
  - For samples of size 15?
  - For samples of size 50?

## Further properties

Any linear combination of independent random variables is also normally distributed.

More generally, the central limit theorem is valid as long as we are sampling many small random events, even if the events have different distributions (as long as no one random event dominates the others).

Why is this cool? It explains why the normal distribution is so common.



Example: Height seems to be determined by a large number of genetic and environmental factors, like nutrition. The “individuals” are genes and environmental factors. Your height is a mean.

# Weibull distributions

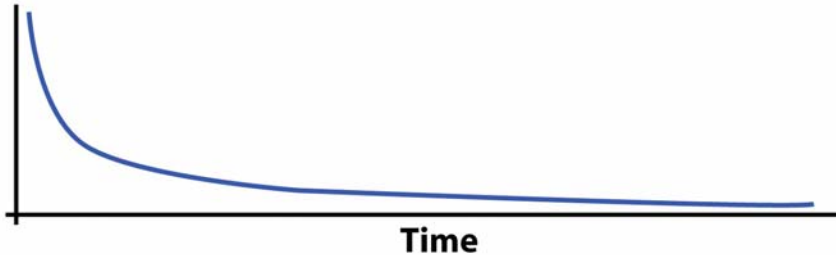
There are many probability distributions beyond the binomial and normal distributions used to model data in various circumstances.

**Weibull distributions** are used to model **time to failure/product lifetime** and are common in engineering to study product reliability.

Product lifetimes can be measured in units of time, distances, or number of cycles for example. Some applications include:

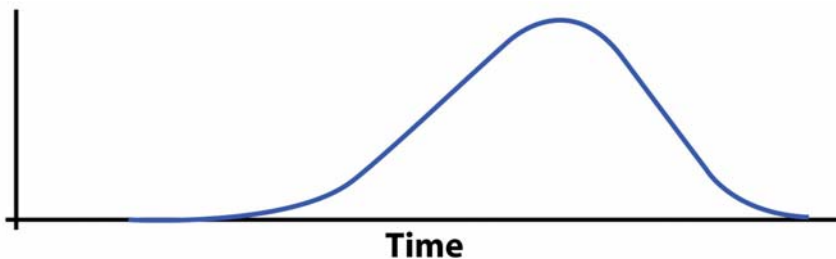
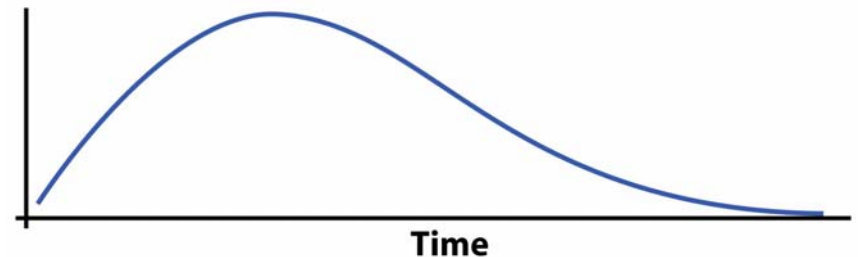
- ❑ Quality control (breaking strength of products and parts, food shelf life)
- ❑ Maintenance planning (scheduled car revision, airplane maintenance)
- ❑ Cost analysis and control (number of returns under warranty, delivery time)
- ❑ Research (materials properties, microbial resistance to treatment)

Density curves of three members of the Weibull family describing a different type of product time to failure in manufacturing:



Infant mortality: Many products fail immediately and the remainder last a long time. Manufacturers only ship the products after inspection.

Early failure: Products usually fail shortly after they are sold. The design or production must be fixed.



Old-age wear out: Most products wear out over time and many fail at about the same age. This should be disclosed to customers.