

## 6 Probability Distributions

### 6.1 How Can We Summarize Possible Outcomes and Their Probabilities?

A **variable** may be **categorical** or **numerical**.

**Definition:** A **random variable** assigns a **number** to each outcome in a population.

Two types of **random variables** are **discrete** and **continuous**.

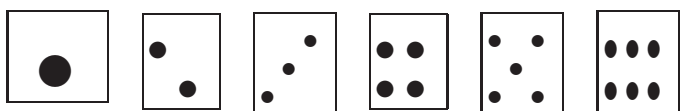
What are some **discrete** random variables?

What are some **continuous** random variables?

#### Discrete distributions

**Example:** (*Discrete case*) Roll a (not necessarily fair) six-sided die once. The possible outcomes are

the **faces** (i.e., dots) on the die.



A different die might have six colors for the six sides  
(like a Rubik's cube).



**Example:** (*Discrete case*) Toss a (not necessarily  
fair) coin once.



**Definition:** The **probability  
distribution** of a **discrete** random variable  $X$   
consists of the possible values of  $X$  along with their  
associated probabilities.

*We sometimes use the terminology **population  
distribution**.*

**Example:** Fair (six-sided) die. Let  $X$  be the numerical outcome. Determine the probability distribution of  $X$ , graph this probability distribution, and compute the mean of the probability distribution.

**Example:** Toss a fair coin 3 times. Let  $X =$  number of heads.

Let  $Y =$  number of matches among the three tosses.

□

**Terminology:** The **expected value** of  $X$  is the **mean** of a random variable  $X$ .

**Example:** (hypothetical) Suppose an airline often overbooks flights, because past experience shows that some passengers fail to show.

Let the random variable  $X$  be the number of passengers who cannot be boarded because there are more passengers than seats.

---

$x$	$P(x)$
0	0.6
1	0.2
2	0.1
3	0.1

---

sum	1
-----	---

---

- (a) Compute the expected value of  $X$ .
- (b) Suppose each unboarded passenger costs the airline \$100 in a ticket voucher. Compute the average cost to the airline per flight in ticket vouchers.

□

## Brief review of means and standard deviations

The **mean**,  $\mu$ , of a random variable is the **average** of all outcomes in the population, and is the limiting value of  $\bar{X}$  as  $n$  gets large.

The **standard deviation**,  $\sigma$ , of a random variable measures the **spread** of all outcomes in the population, and is the limiting value of  $s$  as  $n$  gets large.

The **variance**,  $\sigma^2$ , of a random variable also measures the spread in the population, and is the limiting value of  $s^2$  as  $n$  gets large.

$\sigma$  is more intuitive than  $\sigma^2$ , partly because  $\sigma$  has the same units as the original data.

## Continuous distributions

### Rules for a continuous histogram.

1. The area of a histogram is 1.
2. The **probability** of the random variable taking a value in the interval from “ $a$ ” to “ $b$ ” is the **area** under the probability distribution curve within this interval.
3. The probability distribution function is nonnegative (cannot have negative probability).

**Example:** Let  $X$  be the lifetime of a computer CPU in years, as shown in the graph below. Determine the probability that a new CPU lasts at least 5.5 years.

□

**Example:** Compare means and standard deviations in the graphs below.

## 6.2 How Can We Find Probabilities for Bell-Shaped Distributions?

Here, we focus on the **normal distribution**, which exists in many applications (at least approximately).

Recall the **empirical rule** from section 2.4.

### Empirical Rule

If a large number of observations are sampled from an

approximately normal distribution, then (usually)

1. Approximately 68% of the observations fall within **one** standard deviation,  $\sigma$ , of the mean,  $\mu$ .
2. Approximately 95% of the observations fall within **two** standard deviations,  $\sigma$ , of the mean,  $\mu$ .
3. Approximately 99.7% of the observations fall within **three** standard deviations,  $\sigma$ , of the mean,  $\mu$ .

Suppose  $X$  has a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ .

*Notation:*  $X \sim N(\mu, \sigma)$

$$P(\mu - \sigma < X < \mu + \sigma) = 0.68$$

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.95$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.997$$

**Example:** IQ scores of normal adults on the Weschler test have a symmetric bell-shaped distri-

bution with a mean of 100 and standard deviation of 15.

□

The normal distribution is bell-shaped and symmetric.

## The standard normal distribution

*Notation:*  $Z \sim N(0, 1)$ .

$Z$  represents the number of standard deviations,  $\sigma$ , away from the mean,  $\mu$ .

$Z$  is the “standardized” variable, known as the  $Z$ -score, and has **no units**.

**Example:** Compute  $P(Z < 0)$ ,  $P(Z \leq 0)$ ,  $P(Z > 0)$ , and  $P(Z \geq 0)$ .

□

**Example:** *Using the standard normal table.*

Let  $Z$  be a standard normal random variable.

(a) Determine  $P(Z < 1.26)$ .

(b) Determine  $P(Z > 1.26)$ .

(c) Determine  $P(Z < -1.26)$ .

(d) Determine  $P(Z > -1.26)$ .

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□



Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
∴	∴	∴	∴	∴	∴	∴	∴	∴	∴	∴
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
∴	∴	∴	∴	∴	∴	∴	∴	∴	∴	∴

□

Again consider  $X \sim N(\mu, \sigma)$ .

$$Z = \frac{X - \mu}{\sigma}$$

**Reverse table look-up** uses  $X = \mu + \sigma Z$

$$X \leftrightarrow Z \leftrightarrow \text{probability}$$

**Example:** The length of human pregnancies from conception to birth varies according to a distribution which is approximately normal with mean 266 days and standard deviation 16 days.

(a) Show the empirical rule regarding 95%.

(b) What proportion of pregnancies lasts more than 245 days?

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(c) What proportion of pregnancies last between 245 and 285 days?

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

(d) How long do the longest 20% of pregnancies last?

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

**Example:** Let  $X \sim N(\mu, \sigma)$ . Using the standard normal table, verify the empirical rule regarding 95%. In other words, compute  $P(\mu - 2\sigma < X < \mu + 2\sigma)$  to four significant digits.

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

**Example:** (off the charts)

- (a) Determine  $P(Z < -4)$
- (b) Determine  $P(Z > -4)$
- (c) Determine  $P(Z < 6)$
- (d) Determine  $P(Z < -6)$

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

□

## Quantile-Quantile plots

How do we know if a sample is from a population which is approximately normal?

To construct a Q-Q plot, plot *typical* or *quantile* ordered values from a **normal distribution** against the ordered **observations**.

**Example:** Describe the distributions which likely generated the following Q-Q plots.

□

## 6.3 How Can We Find Probabilities When Each Observation Has Two Possible Outcomes?

### The binomial distribution

**Example:** Toss an unfair coin 5 times, where

$$p = P(\text{heads}) = 0.4.$$

Let  $X$  be the number of heads.

Suppose we want to determine  $P(X = 2)$ .

□

**Factorials:**  $m! = m(m - 1)(m - 2) \cdots 1$  for  
positive integers  $m$ .

**Example:** Compute  $4!$ ,  $5!$ ,  $1!$ , and  $0!$ .

□

A **Bernoulli** trial can have two possible outcomes, *success* or *failure*.

Definition of a **binomial** random variable  $X$ .

1. Let  $n$ , the number of Bernoulli trials, be **fixed** in advance.
2. The Bernoulli trials are **independent**.
3. The probability of success of a Bernoulli trial is  $p$ , which is the same for all observations.

Let  $X$  be the number of *successes*. Then,  $X$  is a binomial( $n, p$ ) random variable.

$$P(X = x) = \frac{n!}{x! (n - x)!} p^x (1 - p)^{n-x}$$

for  $x = 0, 1, 2, \dots, n$

**Example:**  $n = 5$  tosses of an unfair coin.

Assume  $p = P(\text{heads}) = 0.4$  (OR 40% Democrats from a huge population).

Let  $X$  be the number of heads.

Determine the probability distribution of  $X$  and construct the line graph.

$x$	$P(x)$
0	0.0776
1	0.2592
2	0.3456
3	0.2304
4	0.0768
5	0.01024

Determine the probability of obtaining *at least one* heads among the five coin tosses.

In 100 tosses of this coin, on average, how many heads do you expect?

□

**Mean and standard deviation of a**

## binomial random variable

$$\mu = np, \quad \sigma^2 = np(1 - p), \quad \sigma = \sqrt{np(1 - p)}$$

**Example:** *Revisit.* Let  $X \sim \text{Binomial}(n = 5, p = 0.4)$ . Compute the *mean* and *standard deviation* of  $X$ .

□

**Example:** Consider a *huge* population where 30% of the people are Democrats. Let  $X$  be the number of Democrats in a sample of size 1000. Compute the mean and standard deviation of  $X$ .

For large sample sizes (i.e.,  $np \geq 15$  and  $n(1 - p) \geq 15$ ), a *binomial random variable* and a *sample proportion* are approximately **normally distributed** by the **Central Limit Theorem**.

□

**Example:** Viewing the Central Limit Theorem.

- (a) Consider the graphs below for **binomial** random variables, using  $p = 0.3$  and  $n = 1, 2, 3, 4, 5, 10, 15, 20,$  and  $30$ .
- (b) Consider the graphs below for **sampling proportions**,  $\hat{p}$ , using  $p = 0.3$  and  $n = 1, 2, 3, 4, 5, 10, 15, 20,$  and  $30$ .

□

**Example:** *Revisit the Democrats.*

- (a) Use the 95% part of the **empirical rule** on the *binomial random variable*.
- (b) Use the 95% part of the **empirical rule** on the *sample proportion*.

□

## 6.4 How Likely Are the Possible Values of a Statistic? The Sampling Distribution

**Definition:** (Recall) A **statistic** is a quantity computed from a sample.

**Example:**

Recall from section 6.1:

**Definition:** The **probability distribution** of a **discrete** random variable  $X$  consists of the possible values of  $X$  along with their associated probabilities.

**Definition:** The *probability distribution* of a **statistic** is called its **sampling distribution**.

Hence, the **sampling distribution** of a **discrete statistic** consists of the possible values of the *statistic* along with their associated probabilities.

The sampling distribution of a sample proportion,  $\hat{p}$

Recall that a proportion is a special case of a mean, from section 2.3.

**Example:** *Revisit the Democrats.* Sample *independent* observations from a population which is 30% Democrat. Let  $\hat{p}$  be the sample proportion of Democrats.

(a) State the **population distribution** in a chart, and construct the *line graph* of the **population distribution**.

Let  $X = 0$  if non-Democrat, and  $X = 1$  if Democrat.

Note that the *sampling distribution* of  $\hat{p}$  for  $n = 1$  is the same as the *population distribution* of  $X$ .

(b) For  $n = 2$ , state the **sampling distribution** of  $\hat{p}$  in a chart, and construct the *line graph* of the **sampling**

## distribution of $\hat{p}$ .

(c) What happens to the *sampling distribution* of  $\hat{p}$  as the sample size,  $n$ , gets larger?

□

**Example:** *Virginians who exercise.* According to the Centers for Disease Control and Prevention, in 2001, about 48% of Virginian adults achieved the recommended level of physical activity.

*Recommended physical activity is defined as “reported moderate-intensity activities (i.e., brisk walking, bicycling, vacuuming, gardening, or anything else that causes small increases in breathing or heart rate) for at least 30 minutes per day, at least 5 days per week or vigorous-intensity activities (i.e., running, aerobics, heavy yard work, or anything else that causes large increases in breathing or heart rate) for at least 20 minutes*

*per day, at least 3 days per week or both. This can be accomplished through lifestyle activities (i.e., household, transportation, or leisure-time activities).”*

*[http://apps.nccd.cdc.gov/PASurveillance/  
StateSumV.asp?Year=2001](http://apps.nccd.cdc.gov/PASurveillance/StateSumV.asp?Year=2001)*

*[www.cdc.gov/nccdphp/dnpa/physical/stats/  
us\\_physical\\_activity/index.htm](http://www.cdc.gov/nccdphp/dnpa/physical/stats/us_physical_activity/index.htm)*

Take a sample of size  $n = 100$ , and let  $X$  be the number who achieved the recommended level of physical activity. What is the distribution of  $X$ ?

□

Case *A*: Sample **with** replacement. *Hence, observations are independent.*

Case *B*: Sample **without** replacement, but the population size is quite large compared to  $n$ . *Hence, observations are nearly independent.*

*If  $n$  is a small percentage of the population size,*

then sampling **without** replacement is similar to sampling **with** replacement, since sampling the same person more than once would be quite unlikely.

- (a)  $\mu_{\hat{p}} = p$  always.
- (b)  $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$  (called the **standard error** of  $\hat{p}$ ), exactly for Case *A* and approximately for Case *B*.
- (c) (A version of the Central Limit Theorem) The sample proportion  $\hat{p}$  is approximately normal if {rule of thumb}  $np \geq 15$  and  $n(1-p) \geq 15$ , for Cases *A* and *B*.

**Example:** *Revisit Virginians who exercise.* Determine the probability that a majority of Virginians in a sample of size 100 achieve the recommended level of physical activity.

Standard normal table, pp. A1–A2										
<b>z</b>	<b>.00</b>	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
–0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
<b>–0.4</b>	<b>.3446</b>	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
–0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

## Why is the rule of thumb needed?

**Example:** Consider the *sampling distribution* of  $\hat{p}$ , for  $n = 100$  and various  $p$ .

□

## 6.5 How Close Are Sample Means to Population Means?

**Example:** Consider a population consisting of three marbles in an urn, where the marbles are labeled as  $\boxed{2}$ ,  $\boxed{3}$ , and  $\boxed{4}$ . Let  $x$  be the value of a marble drawn.

- (a) Determine the **probability distribution** of  $X$ .
- (b) Graph the *probability distribution* of  $X$ .
- (c) Determine the *mean* of  $X$ .
- (d) Let  $\bar{X}$  be the sample mean, based on **two** observations independently sampled (i.e., **with** replacement) from this population. Determine the **sampling distribution** of  $\bar{X}$ .
- (e) Graph the *sampling distribution* of  $\bar{X}$ .
- (f) Determine the *mean* of  $\bar{X}$ .
- (g) Additional graphs of the *sampling distribution* of  $\bar{X}$  are below, based on independent observations and sample size  $n$ .
- (h) Repeat part (g), using marbles labeled  $\boxed{2}$ ,  $\boxed{3}$ , and  $\boxed{7}$ .

□

Case *A*: Sample **with** replacement. *Hence, observations are independent.*

Case *B*: Sample **without** replacement, but the population size is quite large compared to  $n$ . *Hence, observations are nearly independent.*

(a)  $\mu_{\bar{X}} = \mu$  *always.*

(b)  $\sigma_{\bar{X}} = \sigma/\sqrt{n}$  (called the **standard error** of  $\bar{X}$ ), exactly for Case *A* and approximately for Case *B*.

(c) (A version of the Central Limit Theorem) The sample mean,  $\bar{X}$ , is approximately normally distributed for Cases *A* and *B* (and finite  $\sigma$ ), for **large**  $n$  (usually  $n \geq 30$ , if neither tail of the distribution is too heavy).

(d) (A special case) The sample mean,  $\bar{X}$ , is approximately normally distributed for Cases *A* and *B*

(and finite  $\sigma$ ), if the **original population** is approximately **normally distributed** (for **any** sample size  $n$ ).

**Example:** Suppose  $X \sim N(\mu = 50 \text{ meters}, \sigma = 6 \text{ meters})$ . Sample nine independent observations of  $X$ .

- (a) Determine the *mean* of  $\bar{X}$ .
- (b) Determine the *standard deviation* of  $\bar{X}$ ; i.e., the *standard error* of  $\bar{X}$ .
- (c) Determine the probability that  $\bar{X}$  exceeds 51 meters.

Standard normal table, pp. A1–A2										
z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

□

**Example:** Suppose personal income,  $X$ , in the

U.S. has mean  $\mu = \$40,000$  and standard deviation  $\sigma = \$30,000$ . Sample **without** replacement.

(a) Determine  $P(\bar{X} > \$44,000)$ , for  $n = 64$ .

Standard normal table, pp. A1–A2										
$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
–1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
–1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
–0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

(b) Determine  $P(\bar{X} > \$44,000)$ , for  $n = 100$ .

Standard normal table, pp. A1–A2										
$z$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
–1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
–1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
–1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

(c) What happens to  $P(\bar{X} > \$44,000)$  as we increase  $n$  to 200?

(d) Determine  $P(\bar{X} > \$44,000)$ , for  $n = 10$ .

(e) Determine the 68% part of the empirical rule for

$$n = 100.$$

- (f) Determine the 68% part of the empirical rule for  $n = 10,000$ .

□

## 6.6 How Can We Make Inferences About a Population?

The distribution of the original population is called the **population distribution**.

The distribution of a statistic, such as  $\hat{p}$  or  $\bar{X}$ , is called the **sampling distribution**.

The distribution of one particular data set is called the **data distribution**.

**Example:** *Revisit the Democrats.* Consider a population which is 30% Democrat.

- (a) Graph the **population population**, where a *one* represents a Democrat and a *zero* represents

a non-Democrat.

(b) Let  $\hat{p}$  be the sample proportion of Democrats in a sample of size  $n = 10$ . Graph the **sampling distribution** of  $\hat{p}$ .

(c) In a sample of size 10, suppose that we have four Democrats, three Republicans, and three Independents. Graph the **data distribution**, where a *one* represents a Democrat and a *zero* represents a non-Democrat.

□

**Brief review of formulas** (for independent or nearly independent observations)

*Notation:*  $Z \sim N(0, 1)$

(a)  $Z = \frac{X - \mu}{\sigma}$ , if  $X \sim N(\mu, \sigma)$

(b)  $Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ , if  $\bar{X} \sim N(\mu_{\bar{X}} = \mu, \sigma_{\bar{X}} = \sigma/\sqrt{n})$

Here, we need either the original population to

be approximately normal or a large sample size (usually  $n \geq 30$ , if neither tail of the distribution is too heavy).

Note that  $\sigma_{\bar{X}}$ , the **standard deviation** of  $\bar{X}$ , is also called the **standard error** of  $\bar{X}$ .

$$(c) Z = \frac{\hat{p} - \mu_{\hat{p}}}{\sigma_{\hat{p}}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

Note that  $\sigma_{\hat{p}}$ , the **standard deviation** of  $\hat{p}$ , is also called the **standard error** of  $\hat{p}$ .

Here, we need both  $np \geq 15$  and  $n(1 - p) \geq 15$ .