

# Basic Statistics

Applied Regression and Other Multivariable Methods  
Sections 3-1 – 3-4

2

## Descriptive Statistics

- Numerical summaries designed to describe a feature of a population
- True feature known as a population **parameter**
- Summary based on **sample** known as a **statistic**
- Statistics used to estimate population parameters
- Each statistic can be considered a random variable
  - Observed value based on particular sample
  - Diff sample may result in diff obs value
  - Thus, variation in statistic  $\leftrightarrow$  accuracy
- Measures of center or location
  - Mean
  - Median
- Measures of spread or dispersion
  - Range = Max - Min, IQR =  $Q_3 - Q_1$
  - Variance, Standard Deviation

2-1

## Measures of Center or Location

- **Sample mean** is the arithmetic average
- Denoted as  $\bar{X}$  for underlying variable  $X$

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- $n$  is the sample size
- Can be viewed as point of balance
- $\sum (X_i - \bar{X}) = 0$
- Sensitive to extreme  $X$  values, median less sensitive
- If  $X_i \sim \text{Normal} \rightarrow \bar{X} \sim \text{Normal}$
- Central Limit Thm
  - If  $X_i \sim \text{Other} \rightarrow \bar{X} \approx \text{Normal}$

2-2

## Measures of Spread or Dispersion

- **Deviation** of the  $i$ th observation is  $X_i - \bar{X}$
- Average of deviations in sample is always zero
- "Average" of squared deviations is the **variance**
- Commonly divide sum by  $n - 1$  instead of  $n$

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n - 1}$$

- **Standard deviation**
  - Defined as the square root of the variance
  - Measured in same units as observations

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

- Both sensitive to extreme observations
- Have "nice" properties if  $X \sim \text{Normal}$

2-3

## Random Variables and Distributions

- **Random variable** : cannot predict with certainty
- Many processes affected by chance
  - Genetic trait (random mix of mother and father)
  - Uncontrollable conditions (weather, human interaction)
  - Design of experiment (random allocation of trts, SRS)
  - Lack of complete understanding
- **Probability dist** describes chance of each outcome
- **Discrete Random Variables**
  - A countable number of outcomes  $\leftrightarrow$  "gappy"
  - Dist represented with a histogram
  - Example: Binomial distribution
- **Continuous Random Variables**
  - Take on an uncountable number of values
  - Dist represented with a density curve
  - Example: Normal distribution

2-4

## Binomial Distribution

- Series of  $n$  independent identical trials
- Each trial results in only a success or a failure
- Probability of success,  $\pi$ , is constant
- Independence
  - Can consider each trial separately
  - Combine trials results by multiplying probs
- $X$ =number of successes in  $n$  trials  $\sim \text{Bin}(n, \pi)$ 

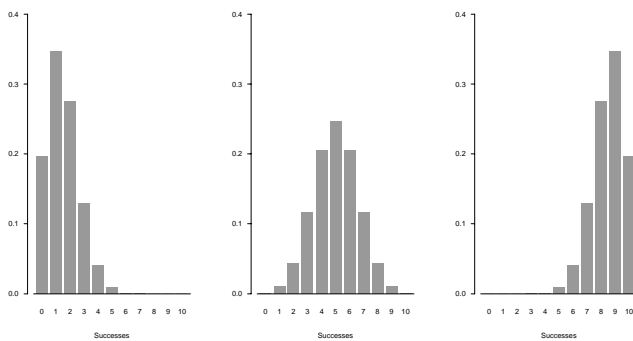
$$\Pr(X = j) = {}_n C_j \pi^j (1 - \pi)^{n-j} \text{ for } 0 \leq j \leq n$$
- ${}_n C_j = \frac{n!}{j!(n-j)!}$  where  $j! = j(j-1)(j-2) \cdots 1$
- Example: Consider  $X \sim \text{Bin}(10, .3)$

$$\Pr(X = 3) = {}_{10} C_3 (.3)^3 (1 - .3)^7 = \frac{10 * 9 * 8}{3 * 2 * 1} (.00222) = 0.266$$

2-5

## Binomial Distribution as Histogram

- Below are three different Binomial Distributions
- Described by  $n = 10$  and  $\pi = \{.15, .5, .85\}$



2-6

## Binomial Distribution Summaries

- Can also describe RV using numerical summaries
- Mean and variance are functions of  $n$  and  $\pi$ 
  - Mean =  $n\pi$
  - Variance =  $n\pi(1 - \pi)$
- Consider three distributions on previous page
  - 1  $n = 10, \pi = .15 \rightarrow \text{Mean} = 1.5, \text{Var} = 1.275$
  - 2  $n = 10, \pi = .5 \rightarrow \text{Mean} = 5, \text{Var} = 2.5$
  - 3  $n = 10, \pi = .85 \rightarrow \text{Mean} = 8.5, \text{Var} = 1.275$

2-7

## Continuous Random Variables

- Can be viewed as extension of discrete RV
- Continuous RV: Two obs can be arbitrarily close
- In other words, # of possible outcomes uncountable
- Cannot talk about  $\Pr(X = x)$  since
  - $\Pr(\text{specific outcome}) \rightarrow 0$  as # of outcomes  $\rightarrow \infty$
- To describe continuous RV use density curve  $f(x)$
- Area under curve equals one
- Define  $\Pr(X \leq x)$  as area under curve to left of  $x$
- $\Pr(X \leq x) = \Pr(X < x)$  since  $\Pr(X = x) = 0$

2-8

## The Normal Distribution

- Denoted  $N(\mu, \sigma)$  where  $\mu = \text{mean}$ ,  $\sigma = \text{std dev}$
- Sometimes denoted with variable subscripts  $N(\mu_X, \sigma_X)$
- Standard Normal density curve ( $\mu = 0, \sigma = 1$ ) is

$$f(z) = Ce^{-.5z^2}$$

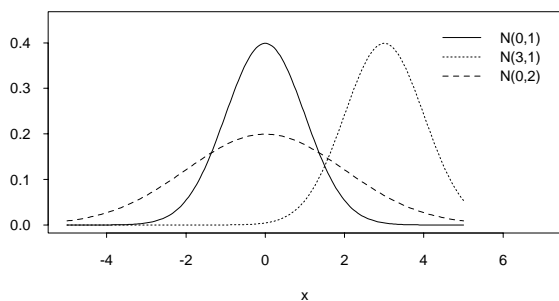
- $e \approx 2.718$  and  $C$  a constant such that area is one
- Features
  - Symmetric and Bell-Shaped
  - Centered at zero ( $\mu$  and median = 0)
  - Standard deviation is one ( $\sigma = 1$ )
  - Satisfies prob features of “nicely shaped” dist
    - $\approx 95\%$  of the observations fall within  $\pm 2$
    - $\approx 68\%$  of the observations fall within  $\pm 1$
    - $\approx 99\%$  of the observations fall within  $\pm 3$
- Computing probabilities under curve
  - Use symmetry and Table A-1

2-9

## Normal Distribution

- Binomial Distribution described by  $n$  and  $\pi$
- Normal distribution described by  $\mu$  and  $\sigma$ 
  - $\mu = \text{mean of the random variable}$
  - $\sigma = \text{standard deviation of random variable}$

$$f(x) = \frac{C}{\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



2-10

## Standardization

- Table A-1 gives probs for std normal ( $\mu = 0, \sigma = 1$ )
- How to compute probabilities for other normal  $X$ ?
- Will use linear transformation :  $Z = aX + b$
- Normal Dist preserved under linear transformation
- Define  $a = 1/\sigma$  and  $b = -\mu/\sigma$

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

- Mean of  $Z$  is  $a\mu + b = \mu/\sigma - \mu/\sigma = 0$
- Std Dev of  $Z$  is  $a\sigma = \sigma/\sigma = 1$
- Subtract off  $\mu$ /divide by  $\sigma$  called **standardization**
- Turns any Normal into standard Normal
- Process used in many statistical procedures

2-11

## Examples

- If the height of this class is normally distributed with mean 70 inches and standard deviation 3 inches, what is the probability that a randomly chosen person is less than 5 foot 8?

5 foot 8 converts to 68 inches

$$\Pr(X < 68) = \Pr(Z < (68 - 70)/3)$$

$$\Pr(Z < -.67) = .2514$$

- If the height of this class is normally distributed with mean 70 inches and standard deviation 3 inches, what is the probability that a randomly chosen person will fall within one standard deviation of the mean?

Within one std dev of the mean is between 67 and 73

$$\Pr(67 < X < 73) = \Pr((67 - 70)/3 < Z < (73 - 70)/3)$$

$$\Pr(-1.0 < Z < 1.0) = .8413 - .1587 = .6826$$

- The specifications for a ball bearing require that its diameter be anything between  $5 \pm .005$  cm. Suppose that past data from the supplier suggests this diameter can be modeled using the normal distribution with  $\mu = 5.002$  and  $\sigma = .002$ . What is the probability that a randomly chosen ball bearing from this supplier will be within specifications?

$$\Pr(4.995 < X < 5.005) =$$

$$\Pr((4.995 - 5.002)/.002 < Z < (5.005 - 5.002)/.002)$$

$$\Pr(-3.5 < Z < 1.5) = .9332 - .0002 = .9330$$

2-12

## Inverse Reading of Table A-1 or pg 714

- Sometimes interest in cutoff for certain %/prob
  - The 95 %-tile of the SAT exam
  - The 90 %-tile for height of 3 mo. boy
- Requires inverse reading of Table A-1
- Examples

What is  $z$  such that  $\Pr(Z < z) = 0.30$ ?

Find 0.30 in table. Nearest is 0.3015  $\rightarrow z = -.52$

What is  $z$  such that  $\Pr(Z > z) = 0.05$ ?

Find 0.95 in table or use page 714.  $z = 1.645$

What is  $x$  so  $\Pr(X > x) = 0.10$  when  $X \sim N(\mu, \sigma)$ ?

Find 0.90 on page 714.  $x = \mu + 1.282\sigma$

2-13

## Normal Approximation of Binomial RV

Consider the experiment of flipping a coin 1000 times and recording the number of heads. Assuming the flips are independent and the coin is fair, the Binomial distribution with  $n = 1000$  and  $\pi = .5$  would be used to answer any probability questions. Suppose we wanted to know the probability that there were no more than 490 heads.

- If we use the Binomial distribution

$$\Pr(X \leq 490) = \sum_{j=0}^{490} {}_{1000}C_j \cdot 5^j (1 - .5)^{1000-j}$$

- Large number of probabilities in sum
- How do we compute  ${}_{1000}C_j$ ?
- Must use computer (answer: .274)

- If we use the Normal Distribution

- Mean is equal to  $n\pi = 500$
- Std Dev is equal to  $\sqrt{n\pi(1-\pi)} = \sqrt{250}$

$$\Pr(X \leq 490) = \Pr\left(Z \leq \frac{490 - 500}{\sqrt{250}}\right) = \Pr(Z \leq -.63) = .2643$$

2-14

## Common Sampling Distributions

- t Distribution** (Table A-2)
  - Used to describe standardized RV ( $\sigma$  unknown)
  - Similar to  $Z$  but depends on degrees of freedom
  - If  $X$  Normal,  $\bar{X}$  and  $S^2$  independent
 
$$\frac{(\bar{X} - \mu)}{S/\sqrt{n}} \sim t_{n-1}$$
  - Often works as approximation for  $T = (\bar{\theta} - \mu_{\bar{\theta}})/S_{\bar{\theta}}$
  - Example: Two-sample t-test w/ constant var
- Chi-square Distribution** (Table A-3)
  - Used to describe sum of squares RVs
  - Consider  $X \sim N(\mu, \sigma)$ 
    - $\sum (X_i - \bar{X})^2 / \sigma^2 \sim \chi_{n-1}^2$
  - Used widely in categorical data analysis
- F distribution** (Table A-4)
  - Used to describe ratio of two indep variances
  - $S_1^2 \sigma_2^2 / S_2^2 \sigma_1^2 \sim F_{n_1-1, n_2-1}$
  - Used extensively in ANOVA
  - Special property:  $t_v^2 \sim F_{1,v}$

2-15