

Exploring Data: Distributions

- Evaluate overall pattern (shape, center, spread) and deviations (outliers).

- Mean (use a calculator):

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum x_i$$

- Standard deviation (use a calculator):

$$s = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

- Median: Arrange all observations from smallest to largest.

- If n is odd, then the median (M) is the middle value.

- If n is even, then the median (M) is the average of the middle two values.

- Quartiles: The first quartile Q_1 is the median of the observations less than the overall median in the ordered list. The third quartile Q_3 is the median of the observations greater than the overall median in the ordered list.

- Interquartile range: $IQR = Q_3 - Q_1$

- Five-number summary:

Minimum, Q_1 , M , Q_3 , Maximum

- Standardized value of x :

$$z = \frac{x - \mu}{\sigma}$$

Sampling Distributions

- Sampling distribution of a sample mean:

○ \bar{x} has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

○ \bar{x} has a Normal distribution if the population distribution is Normal.

○ Central Limit Theorem: \bar{x} is approximately Normal when n is large ($n \geq 30$) even if the population is not normal.

○ Standardized value of \bar{x} :

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Sampling distribution of a sample proportion:

$$\hat{p} = \frac{\text{number of yes}}{\text{total number}}$$

○ \hat{p} has mean p and standard deviation

$$\sqrt{\frac{p(1-p)}{n}}$$

○ \hat{p} is approximately Normal when n is large

○ Standardized value of \hat{p}

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Inference About Proportions

- Standard error: $s.e.(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Large-sample z confidence interval for p :
sample statistic \pm margin of error
sample statistic \pm multiplier \times standard error

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{where } z^* \text{ is from } N(0,1)$$

- z test statistic for $H_0: p = p_0$ if we have a large simple random sample (SRS):

$$z = \frac{\text{Sample statistic} - \text{Null value}}{\text{Standard Error}}$$
$$= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad \text{Get p-values from } N(0,1)$$

Inference About Means

- Standard error: $s.e.(\bar{x}) = \frac{s}{\sqrt{n}}$
- t confidence interval for a population mean if we have a SRS from Normal population:
sample statistic \pm margin of error
sample statistic \pm multiplier \times standard error

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} \quad \text{where } t^* \text{ is from } t(n-1)$$

- t test statistic for $H_0: \mu = \mu_0$ if we have a SRS from a Normal population:

$$t = \frac{\text{Sample statistic} - \text{Null value}}{\text{Standard Error}}$$
$$= \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad \text{Get p-values from } t(n-1)$$

Exploring Data: Relationships

- Evaluate overall pattern (form, direction, strength) and deviations (outliers, influential observations).
- Least-squares regression line (found using computer output): $\hat{y} = b_0 + b_1x$
- Residuals:
 $e_i = \text{residual} = \text{observed } y - \text{predicted } y = y - \hat{y}$

Inference for Regression

- The regression model: We have n observations on x and y . The response y for any fixed x has a Normal distribution with mean given by the true regression line $y = \beta_0 + \beta_1x$ and standard deviation σ . Parameters are β_0, β_1, σ .
- t test statistic for no linear relationship, $H_0: \beta_1 = 0$:

$$t = \frac{\text{Sample statistic} - \text{Null value}}{\text{Standard error}}$$
$$= \frac{b_1 - 0}{s.e.(b_1)} \quad \text{Get p-values from } t(n-2)$$

Where $s.e.(b_1)$ is found using computer output.