

TABLES AND FORMULAS FOR MOORE

Basic Practice of Statistics

Exploring Data: Distributions

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

- Mean (use a calculator):

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

- 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. The median M is located $(n + 1)/2$ observations from the beginning of this list.

- Quartiles: The first quartile Q_1 is the median of the observations whose position in the ordered list is to the left of the location of the overall median. The third quartile Q_3 is the median of the observations to the right of the location of the overall median.

- Five-number summary:

Minimum, Q_1 , M , Q_3 , Maximum

- Standardized value of x :

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

Exploring Data: Relationships

- Look for overall pattern (form, direction, strength) and deviations (outliers, influential observations).

- Correlation (use a calculator):

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

- Least-squares regression line (use a calculator):
 $\hat{y} = a + bx$ with slope $b = rs_y/s_x$ and intercept
 $a = \bar{y} - b\bar{x}$

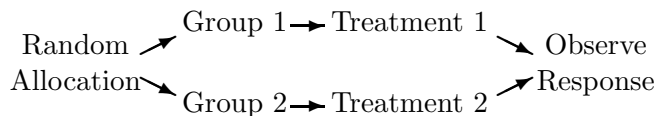
- Residuals:

residual = observed y - predicted $y = y - \hat{y}$

Producing Data

- Simple random sample: Choose an SRS by giving every individual in the population a numerical label and using Table B of random digits to choose the sample.

- Randomized comparative experiments:



Probability and Sampling Distributions

- Probability rules:

- Any probability satisfies $0 \leq P(A) \leq 1$.
- The sample space S has probability $P(S) = 1$.
- If events A and B are disjoint, $P(A \text{ or } B) = P(A) + P(B)$.
- For any event A , $P(A \text{ does not occur}) = 1 - P(A)$

- Sampling distribution of a sample mean:
 - \bar{x} has mean μ and standard deviation σ/\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.

Basics of Inference

- z confidence interval for a population mean (σ known, SRS from Normal population):

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \quad z^* \text{ from } N(0, 1)$$

- Sample size for desired margin of error m :

$$n = \left(\frac{z^* \sigma}{m} \right)^2$$

- z test statistic for $H_0 : \mu = \mu_0$ (σ known, SRS from Normal population):

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \quad P\text{-values from } N(0, 1)$$

Inference About Means

- t confidence interval for a population mean (SRS from Normal population):

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

- t test statistic for $H_0 : \mu = \mu_0$ (SRS from Normal population):

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad P\text{-values from } t(n-1)$$

- Matched pairs: To compare the responses to the two treatments, apply the one-sample t procedures to the observed differences.

- Two-sample t confidence interval for $\mu_1 - \mu_2$ (independent SRSs from Normal populations):

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

with conservative t^* from t with df the smaller of $n_1 - 1$ and $n_2 - 1$ (or use software).

- Two-sample t test statistic for $H_0 : \mu_1 = \mu_2$ (independent SRSs from Normal populations):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

with conservative P -values from t with df the smaller of $n_1 - 1$ and $n_2 - 1$ (or use software).

Inference About Proportions

- Sampling distribution of a sample proportion: when the population and the sample size are both large and p is not close to 0 or 1, \hat{p} is approximately Normal with mean p and standard deviation $\sqrt{p(1-p)/n}$.

- Large-sample z confidence interval for p :

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

Plus four to greatly improve accuracy: use the same formula after adding 2 successes and two failures to the data.

- z test statistic for $H_0 : p = p_0$ (large SRS):

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \quad P\text{-values from } N(0, 1)$$

- Sample size for desired margin of error m :

$$n = \left(\frac{z^*}{m} \right)^2 p^*(1-p^*)$$

where p^* is a guessed value for p or $p^* = 0.5$.

- Large-sample z confidence interval for $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \text{SE} \quad z^* \text{ from } N(0, 1)$$

where the standard error of $\hat{p}_1 - \hat{p}_2$ is

$$\text{SE} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Plus four to greatly improve accuracy: use the same formulas after adding one success and one failure to each sample.

- Two-sample z test statistic for $H_0 : p_1 = p_2$ (large independent SRSs):

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

where \hat{p} is the pooled proportion of successes.

The Chi-Square Test

- Expected count for a cell in a two-way table:

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{\text{table total}}$$

- Chi-square test statistic for testing whether the row and column variables in an $r \times c$ table are unrelated (expected cell counts not too small):

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

with P -values from the chi-square distribution with $\text{df} = (r - 1) \times (c - 1)$.

- Describe the relationship using percents, comparison of observed with expected counts, and terms of X^2 .

Inference for Regression

- Conditions for regression inference: 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 $\mu_y = \alpha + \beta x$ and standard deviation σ . Parameters are α , β , σ .
- Estimate α by the intercept a and β by the slope b of the least-squares line. Estimate σ by the regression standard error:

$$s = \sqrt{\frac{1}{n - 2} \sum \text{residual}^2}$$

Use software for all standard errors in regression.

- t confidence interval for regression slope β :

$$b \pm t^* \text{SE}_b \quad t^* \text{ from } t(n - 2)$$

- t test statistic for no linear relationship, $H_0 : \beta = 0$:

$$t = \frac{b}{\text{SE}_b} \quad P\text{-values from } t(n - 2)$$

- t confidence interval for mean response μ_y when $x = x^*$:

$$\hat{y} \pm t^* \text{SE}_{\hat{\mu}} \quad t^* \text{ from } t(n - 2)$$

- t prediction interval for an individual observation y when $x = x^*$:

$$\hat{y} \pm t^* \text{SE}_{\hat{y}} \quad t^* \text{ from } t(n - 2)$$

One-way Analysis of Variance: Comparing Several Means

- ANOVA F tests whether all of I populations have the same mean, based on independent SRSs from I Normal populations with the same σ . P -values come from the F distribution with $I - 1$ and $N - I$ degrees of freedom, where N is the total observations in all samples.
- Describe the data using the I sample means and standard deviations and side-by-side graphs of the samples.
- The ANOVA F test statistic (use software) is $F = \text{MSG}/\text{MSE}$, where

$$\text{MSG} = \frac{n_1(\bar{x}_1 - \bar{x})^2 + \cdots + n_I(\bar{x}_I - \bar{x})^2}{I - 1}$$

$$\text{MSE} = \frac{(n_1 - 1)s_1^2 + \cdots + (n_I - 1)s_I^2}{N - I}$$

