STA248 Notes

Jenci Wei

Winter 2022

Contents

1 Statistics and Sampling Distributions

Statistic: any quantity whose value can be calculated from sample data

- E.g. $\overline{y}$, s^2

Population parameter: an unknown numerical value

- We are interested to conduct a statistical inference about the population parameter
- E.g. μ, σ^2

Population information

- Size of population: N
- Population mean: μ
- Population variance: σ^2
- Population distribution: Y

Sample information

- Sample size: n
- Samples: $y_1, y_2, \ldots, y_n$
- Sample mean: $\overline{y}$
- Sample variance: s^2
- Mean of the sampling distribution $\overline{y}$: $\mu_{\overline{y}} = E(\overline{y}) = \mu$
- Standard deviation of the sampling distribution $\overline{y}$: $\sigma_{\overline{y}} = \sigma/\sqrt{n}$
  - Called the **standard error** of the mean

Central Limit Theorem

- Refinement of the law of large numbers
- For a large number ($n \geq 30$) of iid RVs $y_1, \ldots, y_n$ with finite variance, the average $\overline{y}$ approximately has a normal distribution, no matter what the distribution of the y_i is
- Let $y_1, \ldots, y_n$ be iid RVs with $E(y_i) = \mu$ and $V(y_i) = \sigma^2 < \infty$. Define

  $$Z_n = \frac{\overline{y} - \mu}{\sigma/\sqrt{n}}$$

  The Z_n follows the standard normal distribution for a large sample size $n \geq 30$, i.e. $Z_n \sim N(0, 1)$ for $n \geq 30$

  - If σ is unknown, then

    $$Z_n = \frac{\overline{y} - \mu}{s/\sqrt{n}} \sim N(0, 1)$$

    where s is the sample SD

The Sampling Distribution of the Sample Proportion

- Consider an event A in the sample space of some experiment with $p = P(A)$. Let y be the number of times A occurs when the experiment is repeated n independent times, and define the sample proportion $\hat{p} = y/n$. Then

1. $E(\hat{p}) = p$
2. $V(\hat{p}) = \frac{p(1-p)}{n}$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$
3. As n increases, the distribution of $\hat{p}$ approaches a normal distribution
   - $\hat{p}$ is approximately normal, provided that $np \geq 10$ and $np(1-p) \geq 10$

Gosset's Theorem

- If $y_1, \ldots, y_n$ is a random sample from a $N(\mu, \sigma)$ distribution, then a RV

$$\frac{\overline{y} - \mu}{s/\sqrt{n}}$$

has the t distribution with $n-1$ degrees of freedom, i.e. t_{n-1}

Chi-Squared Distribution

- Let $y_1, \ldots, y_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then

$$\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - \overline{y})^2$$

has a χ^2 distribution with $n-1$ degrees of freedom (df)

F Distribution

- Let W_1 and W_2 be *independent* χ^2-distributed RVs with ν_1 and ν_2 df, respectively. Then

$$F = \frac{W_1/\nu_1}{W_2/\nu_2}$$

has an F distribution with ν_1 nuumerator degrees of freedom and ν_2 denominator degrees of freedom

2 Point Estimation

An **estimator** is a rule, often expressed as a formula, that tells how to calculate the value of an estimate based on the measurements contained in a sample

- E.g. the sample mean $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is one possible point estimator of the population mean μ

The Bias and Mean Square Error of Point Estimators

- Let $\hat{\theta}$ be a point estimator for a parameter θ. Then $\hat{\theta}$ is an **unbiased estimator** if $E(\hat{\theta}) = \theta$
  - Otherwise $\hat{\theta}$ is **biased**
- The **bias** of a point estimator $\hat{\theta}$ is $B(\hat{\theta}) = E(\hat{\theta}) - \theta$
- The **mean square error** of a point estimator $\hat{\theta}$ is

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}) &= E\left[(\hat{\theta} - \theta)^2\right] \\ &= V(\hat{\theta}) + B(\hat{\theta})^2 \end{aligned}$$

Evaluating the Goodness of a Point Estimator

- The **error** of estimation ϵ is the distance between an estimator and its target parameter, i.e. $\epsilon = |\hat{\theta} - \theta|$

Confidence Intervals

- An **interval estimator** is a rule specifying the method for using the sample measuurements to calculate two numbers that form the endpoints of the interval
  1. We want the interval to contain the target parameter θ
  2. We want the interval to be narrow
- Interval estimators are also called **confidence intervals**
  - The upper and lower endpoints of a confidence interval are called the **upper** and **lower confidence limits**, respectively
  - Suppose that $\hat{\theta}_L$ and $\hat{\theta}_U$ are the (random) lower and upper confidence limits, respectively, for a parameter θ. Then if
  $$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1 - \alpha$$
  the probability $1 - \alpha$ is the **confidence coefficient**

Large-Sample Confidence Intervals

- The endpoints for a $100(1 - \alpha)\%$ confidence interval for θ are given by

$$\begin{aligned} \hat{\theta}_L &= \hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \\ \hat{\theta}_U &= \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}} \end{aligned}$$

Relative Efficiency

- Given two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of a parameter θ. The **efficiency** of $\hat{\theta}_1$ relative to $\hat{\theta}_2$, denoted $\mathrm{eff}(\hat{\theta}_1, \hat{\theta}_2)$, is the ratio
$$\mathrm{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{V(\hat{\theta}_2)}{V(\hat{\theta}_1)}$$

Consistency

- An unbiased estimator $\hat{\theta}_n$ for θ is a **consistent** estimator of θ if

$$\lim_{n\to\infty} V(\hat{\theta}_n) = 0$$

Likelihood Function

- Let $y_1, \ldots, y_n$ be sample observations taken on corresponding RVs $Y_1, \ldots, Y_n$ whose distributions depend on a parameter θ. If $Y_1, \ldots, Y_n$ are discrete RVs, the **likelihood** of the sample, $L(y_1, \ldots, y_n|\theta)$, is defined to be the joint probability of $y_1, \ldots, y_n$.
  - If $Y_1, \ldots, Y_n$ are cts RVs, the likelihood $L(y_1, \ldots, y_n|\theta)$ is the joint density evaluated at $y_1, \ldots, y_n$

The Method of Moments

- Consider the kth moment of a RV, taken about the origin, is

$$\mu'_k = E(Y^k)$$

The corresopnding kth sample moment is the average

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$$

- The method of moments is based on the idea that sample moments should provide good estimates of the corresponding population moments

The Method of Maximum Likelihood

- Suppose that the likelihood function depends on k parameters $\theta_1, \ldots, \theta_k$. Choose the estimates of those parameters that maximize the likelihood $L(y_1, \ldots, y_n|\theta_1, \ldots, \theta_k)$
- The likelihood function is a function of the parameters $\theta_1, \ldots, \theta_k$
  - We sometimes write the lilelihood function as $L(\theta_1, \ldots, \theta_k)$
- Maximum likelihood estimators are referred to as MLEs

3 Statistical Intervals Based On a Simgle Sample

Confidence Interval for Proportion

- Whenever we estimate the SD of a sampling distribution, we call it a **standard error**
- For a sample proportion $\hat{p}$, the standard error is

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- $100(1-\alpha)\%$ confidence interval for the population proportion p is

$$\hat{p} \pm Z_{\alpha/2}\,\mathrm{SE}(\hat{p})$$

  - $100(1-\alpha)\%$ of samples this size will produce confidence intervals that capture the true proportion
  - We are $100(1-\alpha)\%$ confident that the true proportion lies in our interval

- The extend of the interval on either side of $\hat{p}$ is called the **margin of error** (ME):

$$\mathrm{ME} = Z_{\alpha/2}\,\mathrm{SE}(\hat{p})$$

  - $Z_{\alpha/2}$ is called the **critical value** and α is called the **level of significance**

A Confidence Interval for the Mean

- $100(1-\alpha)\%$ confidence interval for the population mean μ:

$$\overline{y} \pm t_{n-1,\frac{\alpha}{2}}\,\mathrm{SE}(\overline{y})$$

  where the standard error of the mean $\mathrm{SE}(\overline{y}) = s/\sqrt{n}$

- If $n \geq 30$, then $100(1-\alpha)\%$ confidence interval for the population mean μ:

$$\overline{y} \pm Z_{\alpha/2}\,\mathrm{SE}(\overline{y})$$

  where the standard error of the mean $\mathrm{SE}(\overline{y}) = s/\sqrt{n}$

Confidence Interval for σ^2

- $100(1-\alpha)\%$ confidence interval for the population variance σ^2:

$$\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right)$$

4 Tests of Hypotheses Based on a Single Sample

Test of Hypothesis

- **Statistical hypothesis**: a statement about the numerical valuue of a population parameter
  - E.g. popuulation mean, population SD
- **Null hypothesis (H_0)**: some claim about the population parameter that the researcher wants to test
  - Either reject or not reject
- **Alternative hypothesis (H_a)**: the values of a population parameter for which the researcher wants to gather evidence to support
  - E.g.

$$H_0 : \mu \leq 24$$
$$H_a : \mu > 24$$

- **Test statistic**: a sample statistic, computed from information provided in the sample
  - Used to decide between the null and alternative hypotheses
- **Type I error**: the researcher rejects the null hypothesis when H_0 is true

$$\alpha = P(\text{Type I error}) = P(\text{Reject } H_0 | H_0)$$

  Value of α is the **level** of the test

- **Rejection region**: the set of possible values of the test statistic for which we could reject H_0
- **Type II error**: the researcher accepts the null hypothesis when H_0 is false

$$\beta = P(\text{Type II error}) = P(\text{Do not reject } H_0 | \neg H_0)$$

- **Observed significance level (p-value)**: the probability, assuming that H_0 is true, of observing a value of the test statistic that is at least as contradictory to the null hypothesis, and supportive of the alternative hypothesis, as the actual one computed from the sample data

Large-Sample α-Level Hypothesis Tests

- $H_0 : \theta = \theta_0$
- $H_a : \begin{cases} \theta > \theta_0 & \text{(upper-tail alternative)} \\ \theta < \theta_0 & \text{(lower-tail alternative)} \\ \theta \neq \theta_0 & \text{(two-tailed alternative)} \end{cases}$
- Test statistic: $Z = \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$
- Rejection region: $\begin{cases} \{z > z_\alpha\} & \text{(upper-tail RR)} \\ \{z < -z_\alpha\} & \text{(lower-tail RR)} \\ \{|z| > z_{\alpha/2}\} & \text{(two-tailed RR)} \end{cases}$

Small-Sample Test for μ

- Assumptions: $Y_1, \ldots, Y_n$ constitute a random sample from a normal distribution with $E(Y_i) = \mu$
- $H_0 : \mu = \mu_0$

- $H_a : \begin{cases} \mu > \mu_0 & \text{(upper-tail alternative)} \\ \mu < \mu_0 & \text{(lower-tail alternative)} \\ \mu \neq \mu_0 & \text{(two-tailed alternative)} \end{cases}$

- Test statistic: $t = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}}$, where $\overline{Y}$ is the sample mean and S is the sample SD

- Rejection region: $\begin{cases} \{t > t_{\alpha,n-1}\} & \text{(upper-tail RR)} \\ \{t < -t_{\alpha,n-1}\} & \text{(lower-tail RR)} \\ \left\{|t| > t_{\alpha/2,n-1}\right\} & \text{(two-tailed RR)} \end{cases}$

Test of Hypothesis Concerning a Population Variance

- Assumptions: $Y_1, \ldots, Y_n$ constitute a random sample from a normal distribution with $E(Y_i) = \mu$ and $V(Y_i) = \sigma^2$

- $H_0 : \sigma^2 = \sigma_0^2$

- $H_a : \begin{cases} \sigma^2 > \sigma_0^2 & \text{(upper-tail alternative)} \\ \sigma^2 < \sigma_0^2 & \text{(lower-tail alternative)} \\ \sigma^2 \neq \sigma_0^2 & \text{(two-tailed alternative)} \end{cases}$

- Test statistic: $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$

- Rejection region: $\begin{cases} \left\{\chi^2 > \chi^2_{\alpha,n-1}\right\} & \text{(upper-tail RR)} \\ \left\{\chi^2 < \chi^2_{1-\alpha,n-1}\right\} & \text{(lower-tail RR)} \\ \left\{\chi^2 > \chi^2_{\alpha/2,n-1} \vee \chi^2 < \chi^2_{1-\alpha/2,n-1}\right\} & \text{(two-tailed RR)} \end{cases}$

Test of Hypothesis: $\sigma_1^2 = \sigma_2^2$

- Assumptions: independent samples from normal populations

- $H_0 : \sigma_1^2 = \sigma_2^2$

- $H_a : \sigma_1^2 > \sigma_2^2$

- Test statistic: $F = \frac{S_1^2}{S_2^2}$

- Rejection region: $F > F_\alpha$, where F_α is chosen so that $P(F > F_\alpha) = \alpha$ when F has $\nu_1 = n_1 - 1$ numerator df and $\nu_2 = n_2 - 1$ denominator df

5 Inferences Based on Two Samples

Comparing two population means: independent sampling – large-sample case

- Properties of the sampling distribution of $\overline{y}_1 - \overline{y}_2$
  1. The mean of the sampling distribution of $\overline{y}_1 - \overline{y}_2$ is $\mu_1 - \mu_2$
     - μ_1 and μ_2 are the means of the two populations
  2. If the two samples are independnet, then the SD of the sampling distribution is

$$\sigma_{\overline{y}_1 - \overline{y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

     - σ_1^2 and σ_2^2 are the variances of the two populations being sampled
     - n_1 and n_2 are the respective sample sizes
     - $\sigma_{\overline{y}_1 - \overline{y}_2}$ is also referred to as the **standard error** of the statistic $\overline{y}_1 - \overline{y}_2$
  3. By the CLT, the sampling distribution of $\overline{y}_1 - \overline{y}_2$ is approximately normal for large samples
- When σ_1^2 and σ_2^2 are known, the $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$(\overline{y}_1 - \overline{y}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- When σ_1^2 and σ_2^2 are unknown, the $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$(\overline{y}_1 - \overline{y}_2) \pm Z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Comparing two population means: independent sampling – small-sample case

- Assumptions
  1. Both sampled populations are approximately normally distributed
  2. The samples have equal population variances (i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$)
  3. Random samples are selected independently of each other
- $100(1-\alpha)\%$ CI for $\mu_1 - \mu_2$ is

$$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

  - S_p^2 is the pooled sample variance where

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$

  - $t_{\alpha/2}$ is based on $n_1 + n_2 - 2$ degrees of freedom

Comparing two population means: independent sampling – hypothesis testing

- Hypotheses
  - $H_0 : \mu_1 - \mu_2 = D_0$

- $H_a : \begin{cases} \mu_1 - \mu_2 > D_0 & \text{(upper-tail alternative)} \\ \mu_1 - \mu_2 < D_0 & \text{(lower-tail alternative)} \\ \mu_1 - \mu_2 \neq D_0 & \text{(two-tailed alternative)} \end{cases}$

- Small-sample case
  - Assumptions
    1. Independent samples
    2. Samples are from normal distribution
    3. $\sigma_1^2 = \sigma_2^2$
  - Test statistic

$$T = \frac{\overline{y}_1 - \overline{y}_2 - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

- Large-sample case
  - Test statistic when σ_1^2 and σ_2^2 are known:

$$Z_c = \frac{\overline{y}_1 - \overline{y}_2 - D_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

  - Test statistic when σ_1^2 and σ_2^2 are unknown:

$$Z_c = \frac{\overline{y}_1 - \overline{y}_2 - D_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$$

Comparing two population proportions: independent sampling

- Properties of the sampling distribution of $\hat{p}_1 - \hat{p}_2$
  1. The mean of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is $p_1 - p_2$, i.e.

$$E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$$

     - p_1 and p_2 are the proportions of the two populations
     - $\hat{p}_1 - \hat{p}_2$ is an unbiased estimator of $p_1 - p_2$
  2. The SD of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

  3. If the sample sizes n_1 and n_2 are large, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal

- Assumptions and conditions when comparing proportions
  1. **Randomization condition**: the data in each group is drawn independently and at random from the target population
  2. **The (least important) 10% condition**: the sample is less than 10% of the population
  3. **Independent group assumption**: the two groups we are comparing are independent of each other
  4. **Success/failure conditions**: both groups are big enough so that at least 10 successes and at least 10 failures have been observed in each group

- In the large-sample case, the $100(1-\alpha)\%$ CI for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Comparing two population proportions: independent sampling – hypothesis testing

- Large-sample test of hypothesis about $p_1 - p_2$: normal statistic:
  - $H_0 : p_1 - p_2 = 0$
  - $H_a : \begin{cases} p_1 - p_2 > 0 & \text{(upper-tail alternative)} \\ p_1 - p_2 < 0 & \text{(lower-tail alternative)} \\ p_1 - p_2 \neq 0 & \text{(two-tailed alternative)} \end{cases}$
  - Test statistic:

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

    where

    $$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Paired Samples and Blocks: Paired t-Test

- Paired data
  - Two results are dependent of each other
  - Since we care about the difference, we could only look at the difference and ignore the original columns
  - Use simple one-sample t-test
  - Sample size is the number of pairs
- Hypotheses
  - We make inferences about the mean of the population of differences, $\mu_d = \mu_1 - \mu_2$
  - $H_0 : \mu_d = d_0$
  - $H_a : \begin{cases} \mu_d > d_0 & \text{(upper-tail alternative)} \\ \mu_d < d_0 & \text{(lower-tail alternative)} \\ \mu_d \neq d_0 & \text{(two-tailed alternative)} \end{cases}$
- Test statistic

  $$t = \frac{\overline{x}_d - d_0}{s_d/\sqrt{n_d}} \sim t_{n_d - 1}$$

  - $\overline{x}_d$ is the sample mean difference
  - s_d is the sample SD of differences
  - n_d is the number of differences (i.e. number of pairs)
  - Assumptions: the population of differences in test scores is approximately normally distributed. The sample differences are randomly selected from the population differences
- Confidence interval: large sample

  $$\overline{x}_d \pm Z_{\alpha/2}\frac{s_d}{\sqrt{n_d}}$$

- Conditions required: a random sample of differences is selected from the target population of differences, and that the sample size n_d is large (i.e. $n_d \geq 30$)

- Confidence interval: small sample

$$\overline{x}_d \pm t_{\alpha/2} \frac{s_d}{\sqrt{n_d}}$$

  - $t_{\alpha/2}$ is based on $n_d - 1$ degrees of freedom
  - Conditions required: a random sample of differences is selected from the target population of differences, and that the population of differences has a distribution that is approximately normal

6 Regression and Correlation

Deterministic Model

- Hypothesizes an exact relationship between variables
- E.g. $y = f(x)$
- Implies that y can always be determined exactly when the value of x is known
- No allowance for error

Probabilistic Model

- Includes both a deterministic component and a random error component
- E.g. $y = f(x) + \text{random error}$

Simple Linear Regression

$$y = \beta_0 + \beta_1 x + \epsilon$$

- The deterministic portion of the model graphs as a straight line
- y is the dependent or response variable
- x is the independent or predictor variable
- $\beta_0 + \beta_1 x$ is the deterministic component
- ϵ is the random error component which is assumed to follow a $N(0, \sigma)$ distribution
- β_0 is the y-intercept of the line
- β_1 is the slope of the line

Estimating Model Parameters

- Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be the observed n-pairs
- The vertical deviation of the point (x_i, y_i) from a line $y = b_0 + b_1 x$ is

$$\text{height of point} \; - \; \text{height of line} = y_i - (b_0 + b_1 x_i)$$

- The sum of squared vertical deviations from the points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to the line is

$$g(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

- The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively, are called the **least squares estimates** whose values minimize $g(b_0, b_1)$
- The estimated regression line or **least squares regression line (LSRL)** is the line whose equation is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

- The least squares estimate of the slope coefficient β_1 of the true regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

- The least squares estimate of the intercept β_0 of the true regression line is

$$b_0 = \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

- Under the normality assumption of the simple linear regression model, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the maximum likelihood estimates

- Notations for sums

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$
$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$
$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Residuals and Estimating σ

- The fitted (or predicted) values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are obtained by successively substituting the x values $x_1, x_2, \ldots, x_n$ into the equation of the LRSL, i.e. the ith fitted value is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = \overline{y} + \hat{\beta}_1 (x_i - \overline{x})$$

- The residuals (estimated error) $e_1, e_2, \ldots, e_n$ are the vertical deviations from the LSRL, i.e. the ith residual is

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right) = (y_i - \overline{y}) - \hat{\beta}_1 (x_i - \overline{x})$$

- The error sum of squares (or residual sum of squares), denoted by SSE, is

$$\text{SSE} = \sum (e_i - \overline{e})^2 = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

- The least squares estimate of σ^2 is

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-2}$$

- The **residual standard deviation** is an estimate of σ given by

$$\hat{\sigma} = \sqrt{\frac{\text{SSE}}{n-2}}$$

- SSE can be computed by

$$\text{SSE} = S_{yy} - \frac{S_{xy}^2}{S_{xx}}$$

Coefficient of Determination

- Total sum of squares: a quantitative measure of the total amount of variation in the observed y values

$$\text{SST} = \sum (y_i - \overline{y})^2 = S_{yy}$$

- The coefficient of determination, denoted by R^2, is given by

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

- R^2 is interpreted as the proportion of observed y variation that can be explained by the simple linear regression model
- The closer R^2 is to 1, the more successful the simple linear regression model is in explaining y variation

Decomposition of Total Sum of Squares

- The total sum of squares can be decomposed by

$$\begin{aligned} \text{SST} &= \sum (y_i - \overline{y})^2 \\ &= \sum \left[(y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}) \right]^2 \\ &= \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \overline{y})^2 \end{aligned}$$

- The regression sum of squares is

$$\text{SSR} = \sum (\hat{y}_i - \overline{y})^2$$

- Therefore

$$\text{SST} = \text{SSR} + \text{SSE}$$

- Coefficient of determination can be rewritten to

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

Inferences About the Regression Coefficient β_1

- Assumptions and conditions
  1. **Linearity assumption**: the *straight enough condition* is satisfied if a scatterplot looks straight
  2. **Independent assumption**: the errors in the true underlying regression model (i.e. the ϵs) must be mutually independent
     - No way of checking whether this holds
  3. **Equal variance assumption**: the variability of y should be about the same for all values of x
  4. **Normal population assumption**: the errors around the idealized regression line at each value of x follows a normal model
     - The response y is normally distributed at any x value
- Properties of the estimated slope
  1. The mean value of $\hat{\beta}_1$ is $E(\hat{\beta}_1) = \beta_1$
     - $\hat{\beta}_1$ is an unbiased estimator of β_1
  2. The variance and SD of $\hat{\beta}_1$ are

$$\begin{aligned} V(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} &= \frac{\sigma^2}{S_{xx}} \\ \sigma_{\hat{\beta}_1} &= \frac{\sigma}{\sqrt{S_{xx}}} \end{aligned}$$

     - σ can be replaced by its estimate $\hat{\sigma}$
  3. The estimator $\hat{\beta}_1$ has a normal distribution
     - Because it is a linear function of independent normal RVs

- As a result, the assumptions of the simple linear regression model imply that

$$T = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t_{n-2}$$

- A $100(1-\alpha)\%$ confidence interval for the slope β_1 of the true regression line is

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2} \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$$

- Hypothesis testing procedures
  - $H_0 : \beta_1 = \beta_{10}$
  - $H_a : \begin{cases} \beta_1 > \beta_{10} & \text{(upper-tail alternative)} \\ \beta_1 < \beta_{10} & \text{(lower-tail alternative)} \\ \beta_1 \neq \beta_{10} & \text{(two-tailed alternative)} \end{cases}$
  - Test statistic:

$$T = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t_{n-2}$$

Inferences for the (Mean) Response

- We want to choose an estimator of the mean y value using the least squares prediction equation

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

  where x^* is some fixed value of x

- Substituting $\hat{\beta}_0$ and $\hat{\beta}_1$:

$$\begin{aligned} \hat{y} &= \sum_{i=1}^{n} \left[\frac{1}{n} + \frac{(x^* - \overline{x})(x_i - \overline{x})}{S_{xx}} \right] y_i \\ &= \sum_{i=1}^{n} d_i y_i \end{aligned}$$

  where $d_i = \left[\frac{1}{n} + \frac{(x^* - \overline{x})(x_i - \overline{x})}{S_{xx}} \right]$

- The coefficients $d_1, \ldots, d_n$ involve the x_is and x^*, all of which are fixed
- Sampling distribution of $\hat{y}$
  1. The mean value of $\hat{y}$ is

$$E[\hat{y}] = E[\hat{\beta}_0 + \hat{\beta}_1 x^*] = E[\beta_0 + \beta_1 x^*] = E[y]$$

  2. The variance of $\hat{y}$ is

$$V(\hat{y}) = \sigma_{\hat{y}}^2 = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right]$$

     The estimated variance of $\hat{y}$ is

$$S_{\hat{y}}^2 = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}} \right]$$

  3. $\hat{y}$ has a normal distribution, because it is a linear function of the y_is, which are normally distributed and independent

- Consequently, the variable

$$t = \frac{\hat{y} - E[y]}{S_{\hat{y}}} \sim t_{n-2}$$

Prediction Interval for a Future Value of y

- Prediction error
  - The prediction error is

  $$\hat{y} - y = \hat{y} - (\beta_0 + \beta_1 x^* + \epsilon)$$

  - The variance of $\hat{y} - y$ is

  $$V[\hat{y} - y] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right]$$

  - The estimated variance of $\hat{y} - y$ is

  $$S^2_{\hat{y}-y} = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(x^* - \overline{x})^2}{S_{xx}}\right]$$

  - Consequently, the variable

  $$t = \frac{(\hat{y} - y)}{S_{\hat{y}-y}} \sim t_{n-2}$$

Correlation

- The **sample correlation coefficient** for the n pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_x}\right) \left(\frac{y_i - \overline{y}}{S_y}\right) = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

- Properties of r
  1. The value of r does not depend on which of the two variables is labelled x and which is labelled y
  2. The value of r is independent of the units in which x and y are measured, i.e. r is unitless
  3. The square of the sample correlation gives the value of the coefficient of determination that would result from fitting the simple linear regression model, i.e. $r^2 = R^2$
  4. $-1 \leq r \leq 1$
  5. $r = \pm 1$ iff all (x_i, y_i) pairs lie on a straight line

7 Analysis of Variance

The **analysis of variance (ANOVA)** is a collection of statistical procedures for the analysis of quantitative responses

- The simplest ANOVA problem is referred to variously as a single-factor, single-classification, or one-way ANOVA and involves the analysis of data samplesd from two or more numerical populations (i.e. distributions)

The **response variable** is the variable of interest to be measured in the experiment

- Also called **dependent variable**
- Typically quantitative

Factors are those variables whose effect on the resonse is of interest to the experimenter

- Also called **independent variables**
- Quantitative factors are measured on a numerical scale

Terminology

- **Factor level**: values of the factor utilized in the experiment
- **Treatment**: factor level combinations utilized in the experiment
- **Experimental unit**: objecto n which the response and factors are observed or measured
- **Design study**: an experiment in which the analyst controls the specification of the treatments and the method of assigning the experimental units to each treatment
- **Observational study**: an experiment in which the analyst simply observes the treatments and the response on a sample of experimental units

Single-Factor ANOVA

- Focuses on comparison of 2 or more populations
- t is the number of populations/treatments being compared
- μ_i is the mean of population i (or the true average resopnse when treatment i is applied)
- The hypotheses of interests are
  - $H_0 : \mu = \mu_2 = \cdots = \mu_t$
  - H_a: at least 2 of the μ_is are different

Sigle-Factor ANOVA Model

- The mathematical model for the data from a **completely randomized design (CRD)** with an unequal number of replicates for each factor level is

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

  where

  - y_{ij} is the resopnse for the jth experimental unit subject to the ith level of the treatment factor, $i \in [1, t]$, $j \in [1, n_i]$
  - n_i is the number of experimental units or replications in ith level of the treatment factor

- The distribution of the experimental errors, ϵ_{ij}, are mutually independent due to the randomizaiton and is assumed to be normally distributed
- τ_i represents the treatment effect
- μ is the overall mean

- We could write the null hpothesis in terms of the treatments effects, where $H_0 : \tau_1 = \tau_2 = \cdots = \tau_t$
- Assumptions
  - The t population or treatment distributions are all normal with the same variance σ^2, i.e. the y_{ij}s are independent and normally distributed with

$$E(y_{ij}) = \mu_i = \mu + \tau_i$$
$$V(y_{ij}) = \sigma^2$$

Single-Factor ANOVA Notations

- The sample means fo the data in the ith level of the treatment factor is represented by

$$\overline{y}_{i.} = \frac{y_i}{n_i}$$

- The **grand mean** is

$$\overline{y}_{..} = \frac{y_{..}}{n}$$

  where

  - $n = \sum_{i=1}^{t} n_i$
  - $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$
  - $y_{..} = \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}$

- A measure of between-samples variation is the **treatment sum of squares (SSTr)**, given by

$$\begin{aligned} \text{SSTr} &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (\overline{y}_{i.} - \overline{y}_{..})^2 \\ &= \sum_{i=1}^{t} n_i (\overline{y}_{i.} - \overline{y}_{..})^2 \\ &= \sum_{i=1}^{t} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{n} \end{aligned}$$

- The **total sum of squares** is

$$\begin{aligned} \text{SSTotal} &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_{..})^2 \\ &= \sum_{i=1}^{t} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{n} \end{aligned}$$

- A measure of within-samples variations is the **error sum of squares (SSE)**, given by

$$\begin{aligned}\text{SSE} &= \sum_{i=1}^{t}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_{i.})^2 \\ &= \text{SSTotal} - \text{SSTr}\end{aligned}$$

Single-Factor ANOVA Result

- When the ANOVA assumptions are satisfied:
    1. SSE and SSTr are independent RVS
    2. $\frac{\text{SSE}}{\sigma^2} \sim \chi^2_{df=n-t}$
    3. When H_0 is true, $\frac{\text{SSTr}}{\sigma^2} \sim \chi^2_{df=t-1}$
- The **mean square for treatments (MSTr)** and the **mean square for error (MSE)** are

$$\text{MSTr} = \frac{\text{SSTr}}{t-1} \qquad\qquad \text{MSE} = \frac{\text{SSE}}{n-t}$$

- When the ANOVA assumptions are satisfied,

$$E(\text{MSE}) = \sigma^2$$

  that is, MSE is an unbiased estimator for σ^2
- Moreover, when H_0 is true,

$$E(\text{MSTr}) = \sigma^2$$

  in this case, MSTr is an unbiased estimator for σ^2
- When ANOVA assumptions are satisfied and H_0 is true, the test statistic $f = \frac{\text{MSTr}}{\text{MSE}}$ has an F distribution with $t-1$ numerator df and $n-t$ denominator df
- Rejection region for level α test: $f > F_{\alpha,t-1,n-t}$
- p-value: area under $F_{t-1,n-t}$ curve to the right of f

Multiple Comparisons in ANOVA

- When H_0 is rejected, we want to know which of the μ_i s are different with each other
- Let $Z_1, Z_2, \ldots, Z_m$ be m independent standard normal RVs, and let W be a χ^2 RV independent of the Z_is. Then the distribution of

$$Q = \frac{\max|Z_i - Z_j|}{\sqrt{W/\nu}} = \frac{\max\limits_{i\in[1,m]} Z_m - \min\limits_{i\in[1,m]} Z_m}{\sqrt{W/\nu}}$$

  is the **studentized range distribution**
- This distribution has 2 parameters
    1. m is the number of Z_is
    2. ν is the denominator df
- We denote the critical value that captures the upper-tail area α under the density curve of Q by $Q_{\alpha,m,\nu}$

Multiple Comparisons in ANOVA Result

- We consider the equal number of replications $n_0 = n_1 = \cdots = n_t$. For each $i < j$, form the interval

$$\overline{y}_{i.} - \overline{y}_{j.} \pm Q_{\alpha,t,n-t}\sqrt{\frac{\text{MSE}}{n_0}}$$

- There are $t(t-1)/2$ such intervals, e.g. $\mu_1 - \mu_2$, $\mu_1 - \mu_3$, etc.
- The simultaneous CI that every interval includes for the correponding value of $\mu_i - \mu_j$ is $100(1-\alpha)\%$

Multiple Comparisons when Sample Sizes are Unequal – Tukey-Kramer Procedure

- Assumption: the t sample sizes $n_1, n_2, \ldots, n_t$ are reasonably close to each other (i.e. *mild imbalance*)
- Let

$$d_{ij} = Q_{\alpha,t,n-t}\sqrt{\frac{\text{MSE}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- Then the probability is approximately $1 - \alpha$ that

$$\overline{y}_{i.} - \overline{y}_{j.} - d_{ij} \le \mu_i - \mu_j \le \overline{y}_{i.} - \overline{y}_{j.} + d_{ij}$$

  for every i and j with $i \neq j$
- The simultaneous confidence level of $100(1-\alpha)\%$ is an approximate

8 Logistic Regression

Logit Function

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = [1+\exp(-\beta_0-\beta_1 x)]^{-1}$$

Odds

- Logistic regression means assuming that $p(x)$ is related to x by the logit function

$$\frac{p(x)}{1-p(x)} = \exp(\beta_0+\beta_1 x)$$

  - The expression on the left side is called the **odds**

Log-Odds

- Taking natural logs on both sides,

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0+\beta_1 x$$

  the logarithm of the odds is a linear function of the predictor

- The slope parameter β_1 is the change in the log-odds associated with a one-unit increase in x
- The quantity e^{β_1} is the **odds ratio**, because it represents the ratio of the odds of success when the predictor variable equals $x+1$ to the odds of success when the predictor variable equals x

Likelihood Function

- There are no analytical solutions for the MLEs $\hat{\beta}_0$ and $\hat{\beta}_1$
- The maximization process must be carried out using iterative numerical methods
- For large n, the MLE has approximately a normal distribution and the standardized variable $\frac{\hat{\beta}_1-\beta_1}{S_{\hat{\beta}_1}}$ has approximately a standard normal distribution

9 Chi-Squared Tests (Extra)

A **multinomial experiment** satisfies the following conditions:

1. The experiment consists of a sequence of n trials for some fixed n
2. Each trial can result in one of the same k possible outcomes (aka categories)
3. The trials are independent
4. Th probability that a trial results in a category i is p_i, which is a constant

The parameters $p_1, \dots, p_k$ must satisfy $p_i \geq 0$ and $\sum p_i = 1$

- Generalization of a binomial experiment, allows each trial to result in one of > 2 possible outcomes

Null hypothesis: p_is are assigned some fixed values, alternative ypothesis: at least one of the p_is has a value different from that asserted by H_0

E.g. an experiment with $n = 50$ and $k = 3$ might yield $N_1 = 22, N_2 = 13, N_3 = 15$

- The N_is are the **observed counts**

$E(N_i) =$ (total number of trials)(hypothesized probability of category i) $= np_{i0}$

- These are the **expected counts** under H_0

Pearson's Chi-Squared Theorem

- When $H_0 : p_1 = p_{10}, \dots, p_k = p_{k0}$ is true, the statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_{i0})^2}{np_{i0}} = \sum_{\text{all categories}} \frac{(\text{observed count - expected count})^2}{\text{expected count}}$$

  has approximately a *chi-squared distribution* with $k - 1$ df

- This approximation is reasonable provided that $np_{i0} \geq 5$ for every i

Chi-Squared Goodness-of-Fit Test

- $H_0 : p_1 = p_{10}, \dots, p_k = p_{k0}$
- H_a : at least one p_i does not equal p_{i0}
- Test statistic value:

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_{i0})^2}{np_{i0}}$$

- Rejection region for level α test: $\left\{\chi^2 \geq \chi^2_{\alpha,k-1}\right\}$

Goodness-of-Fit Tests for Composite Hypotheses

- $H_0 : p_1 = \pi_1(\theta), \dots, p_k = \pi_k(\theta)$ for some $\theta = (\theta_1, \dots, \theta_m)$
- H_a : the hypothesis H_0 is not true

Method of Multinomial Estimation

- Let $n_1, \dots, n_k$ denote the observed values of $N_1, \dots, N_k$. Then $\hat{\theta}_1, \dots, \hat{\theta}_m$ are those values of the θ_js that maximize the expression

$$P(N_1 = n_, \dots, N_k = n_k) \propto [\pi_1(\theta)]^{n_1} \times \cdots \times [\pi_k(\theta)]^{n_k}$$

Fisher's Chi-Squared Theorem

- Under general regularity conditions on $\theta_1, \ldots, \theta_m$, and the $\pi_i(\theta)$s, if $\theta_1, \ldots, \theta_m$ are estimated by maximizing the multinomial expression, then the rv

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - n\hat{P}_i)^2}{n\hat{P}_i} = \sum_{i=1}^{k} \frac{[N_i - n\pi_i(\hat{\theta})]^2}{n\pi_i(\hat{\theta})}$$

  has an approximately a chi-squared distribution with $k-1-m$ df when H_0 is true

- An approximately level α test of H_0 vs. H_a is then to reject H_0 if $\chi^2 \geq \chi^2_{\alpha,k-1-m}$

- This test can be used if $n\pi_i(\hat{\theta}) \geq 5$ for every i

10 Bayesian Estimation (Extra)

Prior Distribution

- A **prior distribution** for a parameter θ, denoted $\pi(\theta)$, is a probability distribution on the set of possible values for θ
- If the possible values of θ form an interval I, then $\pi(\theta)$ is a pdf that must satisfy

$$\int_I \pi(\theta) d\theta = 1$$

- If θ is potentially any value in a discrete set D, then $\pi(\theta)$ is a pmf that must satisfy

$$\sum_{\theta \in D} \pi(\theta) = 1$$

Posterior Distribution

- Suppose $X_1, \ldots, X_n$ have joint pdf $f(x_1, \ldots, x_n; \theta)$ and the unknown parameter θ has been assigned a continuous prior distribution $\pi(\theta)$, then the **posterior distribution** of θ, given the observations $X_1 = x_1, \ldots, X_i = x_i$, is

$$\pi(\theta | x_1, \ldots, x_n) = \frac{\pi(\theta) f(x_1, \ldots, x_n; \theta)}{\int_{-\infty}^{\infty} \pi(\theta) f(x_1, \ldots, x_n; \theta) d\theta}$$

- If $X_1, \ldots, X_n$ is discrete, the joint pdf is replaced by their joint pmf
- Constructing the posterior distribution of a parameter requires a *specific probability model* $f(x_1, \ldots, x_n; \theta)$ for the observed data