1 Numerical Summaries

- Order-based
  - Median - the most middle number in the ordered list
  - Quantiles
    - $Q1$: the most middle number in the left-half ordered list
    - $Q3$: the most middle number in the right-half ordered list
  - Inter-quartile range (IQR) = $Q3 - Q1$
- Average-based
  - mean - the average number of the list

$$\bar{x} = \frac{\sum x_i}{n}$$

  - SD - the rooted square mean of deviation

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

Note that the formula shown before is the "population SD". The "sample SD" denominator is $n-1$.

2 Normal Distribution

Two important R functions for quantiles and percentiles

1. 'qnorm(percentile, mean = 0, sd = 1) $\rightarrow$ quantile'
2. 'pnorm(quantile, mean = 0, sd = 1, lower.tail = TRUE) $\rightarrow$ percentile'

Standardization

$$\frac{X - \mu}{\sigma} \sim N(0, 1)$$

Standard Unit ("Z-score")

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

68%-95%-99.7% Rule

- 68% $\leftrightarrow (-1, 1)$ in "Z score".
- 95% $\leftrightarrow (-2, 2)$ in "Z score".
- 99.7% $\leftrightarrow (-3, 3)$ in "Z score".

3 Correlation Coefficient

Definition

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} = \frac{1}{n}\sum Z_{X_i} \cdot Z_{Y_i}$$

, where the Z score Z_{X_i} and Z_{Y_i} computed by sample deviation.

4 Regression Line

Properties Regression Line $y \sim x (y = b_0 + b_1 x)$

- Line: from $(\bar{x}, \bar{y})$ to $(\bar{x} + SD_x, \bar{y} + rSD_y)$
- Slope: $r\frac{SD_y}{SD_x}$
- Intercept: $\bar{y} - b_1\bar{x}$
- The average residual of the regression line is 0.

Coefficient of determination In value, it equals r^2, where r is the correlation coefficient.

5 Residual Plot

Residual Residual is the number of differences between the prediction and the actual observation.

$$e = y_i - \hat{y_i}$$

6 Probability

Properties and Rules

- $P(Impossible) = 0$ and $P(certain) = 1$
- $P(AB) = P(A|B) \cdot P(B)$
- $P(A \cup B) = P(A) + P(B) - P(AB)$

7 Box Models

Random Draws A random draw X is nothing but a sampling that has size 1. It can be described in math:

$$X = E(X) + [X - E(X)] = E(X) + \epsilon$$

- The first part is the expected value of the random draw, which is equal to the mean in number
- The second part is the standard error of the random draw, which is equal to the SD in number

Sum and Average of Random Draws

- Sum of random draws
  - $E(S) = n \times \mu$
  - $SE(S) = \sqrt{n} \times \sigma$
- Average of random draws
  - $E(\bar{X}) = \mu$
  - $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

8 Unknown Proportion

Prediction Interval

$$P(a \leq static \leq b) = \frac{\gamma}{100}$$

Confidence Interval

$$\bar{x} - mutiplier \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + mutiplier \times \frac{\sigma}{\sqrt{n}}$$

9 One-sample Z-test

Assumption

- The sample is of normal shape (or the sample size is large enough to hold the CLT)
- Each observation in the sample is independent of the others

Z-statistic

$$Z = \frac{\bar{X} - E_0(\bar{X})}{SE_0(\bar{X})}$$

For proportion When we face the discrete data, such as a 01 box, we use "p" to describe the chance that some event happens. Based on this, the $E(\bar{X})$ and $SE(\bar{X})$ can be directly derived as following:

- $E(X) = p$
- $SE(X) = \sqrt{p(1-p)}$

Recalling the knowledge from before, we can fill the number into the formula:

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

- $H_1 : p_0 \neq p$
  - $p = P(Z > |z|) = 2 * P(Z > z) = 2 * pnorm(z, lower.tail = F)$
- $H_1 : p_0 > p$
  - $p = P(Z > z) = pnorm(z, lower.tail = F)$
- $H_1 : p_0 < p$
  - $p = P(Z < z) = pnorm(z)$

then

- $p < \alpha \rightarrow$ reject H_0
- $p > \alpha \rightarrow$ do not reject H_0

For mean

$$Z = \frac{\bar{X} - E_0(\bar{X})}{SE_0(\bar{X})} \sim N(0, 1)$$

Confident Interval version of decision making

$$\bar{x} - multiplier \times \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + multiplier \times \frac{\sigma}{\sqrt{n}}$$

where multiplier $mul = qnorm(1 - ((1-\alpha)/2))$, so if confidence level is 95%, we do calculation $mul = qnorm(0.975)$.

If

- μ_0 fall in the confidence interval $\rightarrow$ do not reject H_0.
- μ_0 does not fall in the confidence interval $\rightarrow$ reject H_0.

10 One-sample T-test

$$T = \frac{\bar{X} - E_0(\bar{X})}{S\hat{E}_0(\bar{X})} \sim t_{n-1}$$

Assumption

- The sample is of normal shape (or the sample size is large enough to hold the CLT)
  - Checked by quantile-quantile plot (QQ plot)
- Each observation in the sample is independent of the others

Critical region of rejection

- $H_1 : \mu \neq \mu_0$
  $(-\infty, -t^*) \cup (t^*, \infty)$ where $t^* = qt(1 - \alpha/2, df = n-1)$
- $H_1 : \mu > \mu_0$
  (t^*, ∞) where $t^* = qt(1 - \alpha, df = n-1)$
- $H_1 : \mu < \mu_0$
  $(-\infty, t^*)$ where $t^* = qt(1 - \alpha, df = n-1)$

Confidence interval

$$(\bar{x} - mul * \frac{\sigma}{\sqrt{n}}, \bar{x} + mul * \frac{\sigma}{\sqrt{n}})$$

where the mul can be computed by $mul = qt(1 - \alpha/2, df = n-1)$.

11 Bootstrap Simulation

Simulation-based P-value Assume we have some data with observed mean μ_{obs} and observed σ_{obs}. The null hypothesis here is $H_0 : p = p_0$.

1. centralize the data, $data - mu_{obs} + p_0$.
2. keep sampling from the data and record the mean value each time
3. compute the p-value
   - $H_A : p \neq p_0 \rightarrow P = mean(abs(recorded\_mean - p_0) >= \mu_{obs})$
   - $H_A : p > p_0 \rightarrow P = mean(recorded\_mean >= \mu_{obs})$
   - $H_A : p \neq p_0 \rightarrow P = mean(recorded\_mean <= \mu_{obs})$

Simulation-based confidence intervals 'quantile(recorded_mean, c(0.025, 0.975))'

12 Two-sample Z-test

Assumption

- Two groups are independent of each other
- Each group follows normal shape (or large enough to apply CLT)

Two-sample Z statistic

$$Z = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim N(0,1)$$

13 Two-sample T-test

Classical Two-Sample T-test

- Assumption:
  - Two groups are independent of each other
  - Each group follows normal shape (or large enough to apply CLT)
  - $\sigma_X = \sigma_Y = \sigma$
- Pooled estimation of common SD σ_p,

$$\hat{\sigma_p} = \sqrt{\frac{\sum(X_i - \bar{X})^2 + \sum(Y_i - \bar{Y})^2}{m+n-2}} = \sqrt{\frac{(m-1)\hat{\sigma}_X^2 + (n-1)\hat{\sigma}_Y^2}{m+n-2}}$$

- Test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\hat{\sigma}_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}$$

Welch's t-test

- Assumption
  - Two groups are independent of each other
  - Each group follows normal shape (or large enough to apply CLT)
- Test statistic

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}} \sim t$$

14 Chi-squared Test

Assumption All expected frequencies E_i are at least 5

Test statistic for goodness of fit

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)}{E_i} \sim \chi^2_{k-1}$$

Test statistic for independence

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{rc-1}$$

P-value

$$p = P(\chi^2 > value_{obs}) = P(\chi^2 > T)$$

15 Inference on Simple Linear Regression

$$Y_i = b_0 + b_1 * x_1 + \epsilon_i$$

Assumption

- The error ϵ_i is independently drawn from an "error box" with mean 0 and SD σ
- The "error box" should be normal-shaped
- Linearity

Test statistic

$$T = \frac{\hat{b_1 - b_1}}{\hat{SE}(\hat{b_1})} \sim t_{n-2}$$

where n is the size of the sample.

Confidence Interval

$$P(\hat{b_1} - u * \hat{SE}(\hat{b_1}) \leq b_1 \leq \hat{b_1} + u * \hat{SE}(\hat{b_1})) = 1 - \alpha$$

16 Multiple Linear Regression

$$Y_i = b_0 + b_1 * x_{1,i} + ... + b_p * x_{p,i} + \epsilon_i$$

Test statistic

$$T = \frac{\hat{b_1 - b_1}}{\hat{SE}(\hat{b_1})} \sim t_{n-(p+1)}$$

where n is the size of the sample, p is number of independent variables.

17 F-test

Hypothesis Example

- H_0: $b_1 = b_2 = b_3 = 0$
- H_1 at least one of the regression coefficient $(b_1, ..., b_p)$ is not zero

Test statistic

$$F \sim F_{p-q, n-(p+1)}$$

where $p - q$ is the number of additional variable between H_0 model and H_1 model, and the p is the number of free variable.

P-value

$$P(F > f) = P(F > t) = pf(t, p-q, n-(p+1), lower.tail = F)$$

- If $p < \alpha \rightarrow$ reject H_0.
- If $p > \alpha \rightarrow$ do not reject H_1.

18 Adjusted R-squared

$$r^2 = 1 - \frac{S\hat{S}E}{S\hat{S}T}$$

$$r^2_{adj} = 1 - \frac{S\hat{S}E/(n-p+1)}{S\hat{S}T/(n-1)}$$

Adjusted R-squared is always smaller than R-squared since we apply penalties to it.

19 Logistic Regression

Odds

$$Odd = \frac{p}{1-p}$$

$$p = \frac{odds}{1 + odds}$$

Logit

$$logit(p_i) = log\frac{p_i}{1-p_i}$$