1 Histogram

Scale

- In **density-scale histogram**
  - The area of each block refers to the proportion, and the total area of the density-scale histogram is 1.
  - The height of each block does not have a specific meaning.
- In **frequency-scale histogram**
  - The area of each block does not have a specific meaning.
  - The height of each block refers to the count, and the sum of all counts of each block is the size of the sample.

Skewness In histogram, we check ”tail” to determine the skewness.

- If one histogram has a tail on the left, then it is left-skewed.
- If one histogram has a tail on the right, then it is right-skewed.

2 Box-plot

Numerical Summaries

- Min and Max
- Quantile ($Q1$, $Q3$)
- Inter-quartile range (IQR) - $Q3 - Q1$
- Mean
- Median
- Lower and Upper threshold $(Q1 - 1.5IQR, Q3 + 1.5IQR)$

Skewness In a box plot, we check the box(IQR) to determine the skewness,

- If one box plot has its box on the left, then it is right-skewed.
- If one box plot has its box on the right, then it is left-skewed.

3 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 $n - 1$.

4 Outliers

It can be detected by the lower and upper thresholds we talked about before.

5 Normal Distribution

Two important R functions for quantiles and percentiles

1. ‘qnorm(percentile, mean = 0, sd = 1) → quantile‘
2. ‘pnorm(quantile, mean = 0, sd = 1, lower.tail = TRUE) → percentile‘

Standardization Any normal distribution can be standardised to N(0, 1). Assume we have a normal distribution X follows $N(\mu, \sigma^2)$, so it can be standardised to

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

Standard Unit (”Z-score”) Also, we have one concept called the standard unit(also known as Z-score), which has a similar formula:

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

68%-95%-99.7% Rule Here is one rule that

- There are 68% of numbers that fall in the scope of (-1, 1) in ”Z score”.
- There are 95% of numbers that fall in the scope of (-2, 2) in ”Z score”.
- There are 99.7% of numbers that fall in the scope of (-3, 3) in ”Z score”.

Application One important application here is that, when we know one dataset follows the normal distribution, we can do an approximation based on the normal distribution.

6 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.

Something should be noted

- Properties
  - This value always fall in the scope $[-1, 1]$.
  - Shifting and Scaling do not change the standard unit. For example, ‘cor(x, y)‘ shows same result with ‘cor(0.2 * x + 3, 3 * y - 1)‘.
  - Interchanging the variable does not affect the correlation coefficient. For example, ‘cor(x, y)‘ shows same result with ‘cor(y, x)‘
- Warning
  - $r = 0.8$ does not mean 80% of the points tightly around the line.
  - Outliers can overly influence the correlation coefficient.
  - The correlation coefficient can only detect the linear association; nonlinear association cannot be detected by it.
  - The same correlation coefficient can be raised from very different data.

7 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 The coefficient of determination is a metric that measures how well the performance of the regression line works on this data. It describes how many dependent variables can be explained by this model compared to the baseline model. In value, it equals r^2, where r is the correlation coefficient.

8 Residual Plot

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

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

Application Residual plot can be used as a diagnostic plot when checking whether the linear fit is appropriate for the data.

- If it shows no pattern, then it is appropriate.
- If it shows a fanning pattern or quadratic pattern, then it is not.

9 Probability

Properties and Rules

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

Chance Simulation Sometimes we do sampling from data set, say, we draw a ticket from a black box. There are two ways of sampling, one is sampling with replacement and another is sampling without replacement.

- When we perform the sample with replacement, we put back the ticket after each single draw. So, each time we put our hands in that black box, the status in the box is the same. Hence, each single draw is independent of the others.
- When we perform the sample without replacement, we do not put the ticket back. Strictly speaking, each single draw is not independent, because each time we put our hands in the box, the status inside changes. However, in some cases, in a very large sampling process (sample size way smaller than the population size, < 10%), we can consider that the sample is almost independent.

10 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
  - $E(\bar{X}) = \mu$
  - $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

11 Central Limit Theorem (CLT)

CLT is nothing but a theorem that says that the sample sum and sample average will follow the normal shape when the sample size is ”large enough”.

- When the sample is reasonably symmetric and does not have too many outliers, $n = 5$ or 10 should be considered as ”large enough”
- When the sample is very skewed, maybe $n > 100$ should be considered as ”large enough”.

12 Unknown Proportion

Prediction Interval If we know the data follows a normal shape, we can calculate the prediction interval. That means the sample sum(or average) has γ% of chance of falling in the interval $[a, b]$.

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

Confidence Interval The idea of the confidence interval is that when we do not know the actual value of p(or μ), we can inversely use the observed value (based on the prediction interval) to make a prediction on it.

For example, the 95% prediction interval of the sample mean is

$$[E(\bar{X}) - mutiplier \times SE(\bar{X}), E(\bar{X}) + multiplier \times SE(\bar{X})]$$

with $E(\bar{X}) = \mu$, $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$ and μ known, $\bar{x}$, σ unknown, we can do following derivation:

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

13 One-sample Z-test

The key idea of the one-sample Z-test is that we make a hypothesis, and then make a mock distribution based on that hypothesis. After that, we do an observation. If the observation value is far away from the mock distribution we built, then this distribution (and its hypothesis) must be fake.

False alarm rate (or level of significance) is a concept that describes the consistency between data and the hypothesis.

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:

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

Then we calculate the P-value here, which means that "the probability of observing something more extreme than the observed sample when the hypothesis is true".

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

So, when the p-value is less than the significance level, it means that this kind of extreme value has a very limited chance of occurring. Only in that case, we reject the null hypothesis. Otherwise, we do not have sufficient evidence to reject the null hypothesis.

For mean When we face the continuous data, we directly use $E_0(\bar{X})$ and $SE_0(\bar{X})$ to perform z-test. All the processing is the same as the "for proportion" version, but uses the formula

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

Confident Interval version of decision making In the previous version, we compared the p-value and significance level. But with the help of the confidence interval, we can make the same decision.

With the known significant level α, we can compute the multiplier (also called the critical value). Recall the definition of confidence interval, we know that

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

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

If we find that

- the hypothesis does not fall in the confidence interval, then we should reject it.
- the hypothesis falls in the confidence interval, then we do not reject it.

14 One-sample T-test

The t-test is a test that we apply when we do not know the exact value of SD. In that case, we replace the $SE_0(\bar{X})$ with $\hat{SE}_0(\bar{X})$, the little hat here means the estimated value from the sample.

Hence, the test statistic becomes to

$$T = \frac{\bar{X} - E_0(\bar{X})}{\hat{SE}_0(\bar{X})}$$

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

Student's t distribution Due to the change in formula, the t-statistic does not follow the normal distribution but Student's t distribution with degree of freedom of $n-1$, where n is the size of the sample.

Critical region of rejection is a concept related to the test statistic. If the test statistic falls in the scope of the region of rejection, then we reject the null hypothesis. Otherwise, we do not reject the null hypothesis.

The critical region of rejection can be computed by the significance level α, if we have the alternative hypothesis

- $H_1 : \mu \neq \mu_0$
  - let $t^* = qt(1-\alpha/2, df = n-1)$, then the region of rejection is $(-\infty, -t^*) \cup (t^*, \infty)$
- $H_1 : \mu > \mu_0$
  - let $t^* = qt(1-\alpha, df = n-1)$, then the region of rejection is (t^*, ∞)
- $H_1 : \mu < \mu_0$
  - let $t^* = qt(1-\alpha, df = n-1)$, then the region of rejection is $(-\infty, t^*)$

Confidence interval for estimating the population mean with unknown population SD. It is quite similar to the version of the Z-test. The population mean falls within the scope of

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

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

15 Bootstrap Simulation

The key idea of bootstrap simulation is to sample from the observations to simulate we have a population. Based on that, we built a distribution of statistics.

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

Simulation-based confidence intervals We can call ‘quantile(recorded_mean, c(0.025, 0.975))‘ when the significant level is 5%.

16 Two-sample Z-test

Two-box model Two-box model is the combination of two box models. In the two-box model, we have two groups

- The first group as a random sample $X_1, ..., X_m$ taken (with repl.) from a box with
  - mean μ_X
  - SD σ_X
- The second group as a random sample $Y_1, ..., Y_m$ taken (with repl.) from a box with
  - mean μ_Y
  - SD σ_Y

Based on the proprieties of $E()$ and $SE()$, we can do following derivation

- $E(\bar{X} - \bar{Y}) = \mu_X - \mu_Y$
- $SE(\bar{X} - \bar{Y}) = \sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}$

Assumption

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

Two-sample Z statistic

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

17 Two-sample T-test

Same as the one-sample z-test and one-sample t-test, when we do not know the exact value of σ_X and σ_Y, we use estimated value $\hat{\sigma_X}$ and $\hat{\sigma_Y}$ to replace it.

Follow the previous formula, the different assumption here lead to the different t-test.

Classical Two-Sample T-test Classical Two-sample t-test has following assumption:

- Two groups are independent of each other
- Each group follows normal shape (or large enough to apply CLT)
- $\sigma_X = \sigma_Y = \sigma$

It uses pooled estimation to calculate the common SD σ_p,

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

Based on that, we have statistics

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

Welch's t-test Welch's t-test does not assume two groups have same σ, so it is flexible on the assumption:

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

It still follows students' t distribution, but has difference in the degree of freedom calculation with the classical two-sample t-test. The degree of freedom can be calculated using R code.

18 Chi-squared Test

The chi-squared test is a kind of test that tests whether the observation is aligned with the expectations.

Assumption The χ^2 distribution is a "large-sample approximation" to the exact sampling distribution of Pearson's statistic when H_0 is true. We can simply check whether all expected frequencies E_i are at least 5 to guarantee this assumption (which is also called the "rule of thumb").

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 Does not like the calculation process in Z-test and t-test, even though we do make a two-side test here, the p-value is still deduced by one-tail.

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

where T is the test statistic we calculated before.

19 Inference on Simple Linear Regression

Probability view of Simple Linear Regression If we take each prediction as one single random draw, like we decompose the single draw to the mean and error part, we can decompose each prediction into two parts.

$$Y_i = b_0 + b_1 \times x_i + \epsilon_i$$

If we apply the same t-test here, we can test whether x_i has a linear relationship with Y_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

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

where n is the size of the sample.

Confidence Interval This is the same as the part we calculate the confidence interval fot t-test.

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

20 Multiple Linear Regression

Multiple Linear Regression is the updated version of simple linear regression. Essentially, there is not much difference between them. The only difference occurs in two places:

- The first one is that the fitted model is changed to

$$Y_i = b_0 + b_1 * x_1 + ... + b_p * x_p$$

- The second one is that the test statistic no longer follows t_{n-2} but $t_{n-(p+1)}$

21 F-test

We used a t-test in the previous section to test whether one independent variable is significant in explaining the dependent variable. There is one more appropriate way to do this, which is the F-test.

Hypothesis Example

- H_0: $b_1 = b_2 = b_3 = 0$
  - Which means none of the independent variables help explain the dependent variable
  - The corresponding null model

$$Y_i = b_0 + \epsilon_i$$

- H_1 at least one of the regression coefficient $(b_1, ..., b_p)$ is not zero
  - At least one of the independent variables affects the explanation of the dependent variable
  - The corresponding alternative model

$$Y_i = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + \epsilon_i$$

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 $p < \alpha$, then it means those variables have a significant impact on explaining the independent variable. We reject H_0.
- If $p > \alpha$, then it means those variables do not have a significant impact on explaining the independent variable. We do not reject H_1.

22 Adjusted R-squared

The normal r^2 does not penalties the variables that are used to explain the independent variable, which means more variables used, the higher the r^2 value. That does not make sense because it will lead to many variables that have a limited impact on explaining the independent variable. That is one kind of overfitting.

So we are trying to impose penalties on the number of variables used in the model. That leads to adjusted r_2.

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

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

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

23 Model Selection

When we know the F-test, we can do the model selection based on which variables show more significance in explaining the independent variable.

There are two ways of model selection,

- The first one is "backward model selection"
  - Start from the full model, a model with all of the variables
  - Then do the F-test to calculate the less significant variable (threshold: 5 in F test statistic) and remove it
  - Keep doing the last two steps until all variable has F-test statistic value over 5.
- The second one is "forward model selection"
  - Start from the null model, a model only with an intercept
  - Then do the F-test to calculate the most significant variable (threshold: 5 in F test statistic) and add it
  - Keep doing the last two steps until all unadded variable has the F-test statistic value less than 5.

24 Logistic Regression

Logistic Regression is a multiple linear regression that predicts a probability rather than a number.

Due to the feature of probability, we do not directly predict the chance. We convert it to its odds.

Odds

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

- when $odd > 1$, then the event is more likely to happen than not
- when $odd = 1$, then the event is equally likely to happen than not $(50\% - 50\%)$
- when $odd < 1$, then the event is less likely to happen than not

Also, we can convert odd to p inversely by

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

Logit

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