Topic 02 Exploring Data

Sample mean = sum of data / size of data.

Given a data point x_i, its deviation from the sample mean is $D_i = x_i - \overline{x}$.

Sample mean balances the absolute deviations: $\sum_{x_i<\overline{x}}|x_i - \overline{x}| = \sum_{x_i>\overline{x}}|x_i - \overline{x}|$.

Sample median is the middle data point ($\widetilde{x}$).

The sample median is the half way point on the histogram.

The sample median is robust (健壮) and is a good summary for skewed (倾斜) data as it is not affected by outliers.

left skewed data: $\overline{x} < \widetilde{x}$, right skewed data: $\overline{x} > \widetilde{x}$, symmetric data: $\overline{x} = \widetilde{x}$.

Root mean square: $RMS = \sqrt{sample\ mean(numbers^2)}$.

Population Standard Deviation: $SD_{pop} = RMS\ of\ deviations = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 / n}$.

Sample Standard Deviation: $SD_{sample} = \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 / n-1}$.

Variance: $Var_{pop} = {SD_{pop}}^2$, $Var_{sample} = {SD_{sample}}^2$.

Another formula: $Var(X) = Mean(X^2) - Mean(X)^2$

Standard units (Z score) of a data point = how many standard deviations is it below or above the mean: $Z_i = {x_i - \overline{x}}/{SD}$.

IQR is range of middle 50% data. Q_1 is the 25-th percentile (1st quartile) and Q_3 is the 75-th percentile (3rd quartile). $\widetilde{x} = Q_2$. $IQR = Q_3 - Q_1$. IQR is robust.

Lower thresholds: $LT = Q_1 - 1.5IQR$, upper thresholds: $UT = Q_3 + 1.5IQR$.

Topic 03 Normal Curve

General Normal Curve (X) is denoted by N(mean, Variance) or N(μ, σ^2).

Standard Normal Curve (Z) is denoted by N(0, 1).

pnorm(x) gives the lower tail area P(Z<x). pnorm(x, m, sd, lower.tail=F) gives upper tail area of P(X>x), X is N(μ, σ^2).

68 95 99.7 rule: $P(\mu - \{1|2|3\}\sigma \le X \le \mu + \{1|2|3\}\sigma) \approx \{68|95|99.7\}\%$

Rescaling: X following N(μ, σ^2), $P(X < a) = P(Z < {a - \mu}/{\sigma})$

Symmetric: P(Z < -a) = P(Z > a), P(X < μ - a) = P(X > μ + a)

Topic 04 Correlation and Linear Model

Bivariate data involves a pair of variables (x_i, y_i).

Bivariate data can be summarized by five numerical summaries: $(\bar{x}, SD_x)$, $(\bar{y}, SD_y)$ and correlation coefficient (r).

$$r = \sum_{i=1}^{n}(x_i - \bar{x})\,(y_i - \bar{y}) \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

r→0: no linear dependency. r→±1: cluster around the line.

Positive r: the cloud slopes up. Negative r: the cloud slopes down.

Shift and scale invariant: r(x, y) = r(ax + b, cy + d). Symmetry: r(x, y) = r(y, x).

Outliers can overly influence the correlation coefficient.

Baseline prediction: $\hat{y}_i = \bar{y}$.

Regression (回归) line connects $(\bar{x},\ \bar{y})$ to $(\bar{x} + SD_x,\ \bar{y} + r \cdot SD_y)$.

Prediction: $\hat{y}_i = a + b \cdot x_i$. Slope(b): $r \cdot {SD_y}/{SD_x}$. Intercept(a): $\bar{y} - b \cdot \bar{x}$.

To predict x using y, we need to refit the model. $(\bar{y},\ \bar{x})$ to $(\bar{y} + SD_y,\ \bar{x} + r \cdot SD_x)$.

A residual (prediction error)(残差) is the vertical distance of a point above or below the regression line.$e_i(a, b) = y_i - \hat{y}_i = y_i - (a + b \cdot x_i)$.

Sum and mean of residuals are zero:
$\sum_{i=1}^{n} e_i(a, b) = \sum_{i=1}^{n}(y_i - \bar{y}) - b\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Regression line minimizes the sum of squares of the residuals.

Sum of squared residuals (or SSE for sum of squared errors):
$SSE = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$.

Sum of squared deviations about sample mean (or SST for sum of squared total): $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

$SST \geq SSE \geq 0,\ r^2 = 1 - {SSE}/{SST} = 1 - {SD(e)^2}/{SD(y)^2}$,

$SSE = (n-1)SD(e)^2,\ SST = (n-1)SD(y)^2$.

Coefficient of determination (r^2) gives the proportion of variation in the dependent variable y that can be explained by the model.

The higher the value of r^2, the more successful is the simple linear regression model in

explaining y variation.

A residual plot graphs the residuals vs x. If the linear fit is appropriate for the data, it should show no pattern.

Topic 05 Sampling Data

Probability: the percentage of time a certain event is expected to happen, if the same process is repeated long-term (infinitely often).

P(Event) = 1 – P(Complement Event)

Conditional Event: the chance of Event A occurs given that Event B has occurred. P(Event A | Event B)

P(Event A and Event B) = P(Event A) × P(Event B | Event A)

P(Event A or Event B) = P(Event A) + P(Event B) - P(Event A and Event B)

Mutually exclusive: the occurrence of one event prevents the occurrence of the other.

Independence: when A and B satisfy P(Event A | Event B) = P(Event A).

sample(1:6, m, rep=T) simulates a box model. In a box model, there are N tickets in a box, and we want to draw m tickets from the box.

Topic 06 The Box Model

Given $y_i = ax_i + b\ (a \neq 0)$, we can get population mean: $\bar{y} = a\bar{x} + b$ and SD: $SD_{pop}(y) = |a| \cdot SD_{pop}(x)$.

The box model is a collection of N objects (tickets). Box is a population.

We can take a random sample of a certain size n from the box (with or without replacement). A random draw is a random sample with n=1.

Expected value of a random draw: mean of the box, E(X).

Standard error of a random draw: SD of the box, SE(X).

Random draw = Expected value + Chance error: $X = E(X) + X - E(X) = E(X) + \varepsilon$.

Chance error ε is a random draw from an error box (deviation box having mean 0).

Because error box has mean 0, the standard error is also the RMS of the error box: $SE(X) = SD(box) = RMS(deviation) = RMS(\varepsilon) = RMS(\varepsilon - 0) = SD(\varepsilon)$.

Expected value of sum is sum of expected values: $E(X + Y) = E(X) + E(Y)$.

Squared SE of the sum is the sum of the squared SEs: $SE(X + Y)^2 = E(X)^2 + E(Y)^2$.

Sum of draws: $E(X_1 + \cdots + X_n) = n \cdot E(X)$, $SE(X_1 + \cdots + X_n)^2 = n \cdot SE(X)^2$.

Mean of draws: $E(\bar{X}) = E\left(\sum_{i=1}^{n} X_i / n\right) = n \cdot E(X) / n = E(X)$,

$SE(\bar{X}) = \sqrt{n \cdot SE(X)^2} / n = SE(X) / \sqrt{n}$.

Topic 07 Central Limit Theorem

$P(Z < z)$ is often called the CDF of "standard normal" denoted by $\Phi(z)$.

If $S = X_1 + \cdots + X_n$ is the sum of random sample (with replacement) of size n from a box with mean μ and SD σ, $\bar{X}$ is the mean of random sample ($\bar{X} = S/n$), then for large n:

$S \sim N(n\mu, (\sigma\sqrt{n})^2)$, $\bar{X} \sim N(\mu, (\sigma/\sqrt{n})^2)$.

That is to say: $P(S \le s) = P(\frac{S-n\mu}{\sigma\sqrt{n}} \le \frac{s-n\mu}{\sigma\sqrt{n}}) \approx \Phi(\frac{s-n\mu}{\sigma\sqrt{n}})$, and

$P(\bar{X} \le x) = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{x-\mu}{\sigma/\sqrt{n}}\right) \approx \Phi\left(\frac{x-\mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{s-n\mu}{\sigma\sqrt{n}}\right)$.

Topic 08 Unknown Properties

0-1 box: a special box only contains 0 and 1.

Let p (0 <= p <= 1) denote the proportion of 1s in the box, and N be the size of the box.

Then, the box contains $(1-p)N$ 0s and pN 1s.

Mean: $\mu = pN/N = p$, SD: $\sigma = \sqrt{mean.sq. - (mean)^2} = \sqrt{p(1-p)}$.

Take n draws, $E(S) = n\mu$, $SE(S) = \sigma\sqrt{n}$, $E(\bar{X}) = \mu$, $SE(\bar{X}) = \sigma/\sqrt{n}$.

Interval Prediction: A γ% chance that S lands in [a, b]: $P(a \le S \le b) = \gamma\%$, or a γ% chance that $\bar{X}$ lands in [c, d]: $P(c \le \bar{X} \le d) = \gamma\%$. (The purpose is to calculate abcd using γ) (ab and cd are symmetry)

[a, b] is a γ% confidence interval for S. [c, d] is a γ% confidence interval for $\bar{X}$.

Applying the Central Limit Theorem:

$P(a \le S \le b) = P\left(\frac{a-n\mu}{\sigma\sqrt{n}} \le \frac{S-n\mu}{\sigma\sqrt{n}} \le \frac{b-n\mu}{\sigma\sqrt{n}}\right) \approx \Phi\left(\frac{b-n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a-n\mu}{\sigma\sqrt{n}}\right)$,

$P(c \le \bar{X} \le d) = P\left(\frac{c-\mu}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{d-\mu}{\sigma/\sqrt{n}}\right) \approx \Phi\left(\frac{d-\mu}{\sigma/\sqrt{n}}\right) - \Phi(\frac{c-\mu}{\sigma/\sqrt{n}})$.

Therefore: $a = n\mu - \alpha\sigma\sqrt{n}$, $b = n\mu + \alpha\sigma\sqrt{n}$,

$c = \mu - \beta \cdot \sigma/\sqrt{n}$, $d = \mu + \beta \cdot \sigma/\sqrt{n}$. α and β are calculated by qnorm, they are

actually z score.

For 0-1 box, $E(\bar{X}) = \mu = p,\ SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$.

Therefore, $c = p - \beta\sqrt{\frac{p(1-p)}{n}},\ d = p + \beta\sqrt{\frac{p(1-p)}{n}}$.

Consistency: with γ% chance, sample means fall into the prediction interval around p. Those sample means in the interval are considered consistent with the parameter p.

If p is unknown, change the inequality to: $-\beta \le \frac{\bar{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \le \beta$. Solve this inequality and we will get a confidence interval (confidence γ%) for unknown p.

Confidence interval means γ% of the time, the interval covers the true p value.

Topic09 Z test

If the observed $\bar{X}$ is within the range (c, d) we would conclude "data is consistent with the hypothesis p value";

If the observed $\bar{X}$ is outside the range (c, d) we would "reject" the hypothesis p value.

False alarm rate (or level of significance): the chance we reject the hypothesis when it is true.

Z statistic measure how many SEs away the observed value $\bar{X}$ is from the expected value, converting the observed $\bar{X}$ into standard units, assuming the hypothesis is true. $z = \frac{\bar{X}-E_0(\bar{X})}{SE_0(\bar{X})}$, E_0 and SE_0 are computed assuming the hypothesis is true. (α and β are Z statistic)

Use z score to calculate p value: $p = P(Z < -|z|) + P(Z > |z|) = P(2 \cdot Z > |z|) = 2 * pnorm(abs(z), lower.tail = F)$.

Hypothesis test H_0: $p = p_0$ (the unknown proportion p is equal to the special value p_0).

Null hypothesis is H_0: $p = p_0$. Alternative hypothesis (double sided test) is H_1: $p != p_0$.

Rejecting H_0 ($p = p_0$): Reject at (100-γ)% level of significance if and only if $\bar{X}$ is NOT in the γ% prediction interval for p_0, that is if: $\bar{X} < p_0 - z_0\sqrt{\frac{p_0(1-p_0)}{n}}$ or $\bar{X} > p_0 + z_0\sqrt{\frac{p_0(1-p_0)}{n}}$; equivalently if $|z| = \frac{|\bar{X}-p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_0$. ($z_0$ is the one given by confidence γ%, also called multiplier / critical value, z_0 = qnorm((1 - γ%) / 2)

Consistent with H_0 ($p = p_0$): If $\bar{X}$ lands within the prediction interval, i.e. if:

$|z| = \frac{|\bar{X} - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leq z_0$, we say the data is consistent with H_0 at the (100-γ)% level of significance. (we do not accept H_0, just keep it)

γ% is called confidence level; (100-γ)% is called significance level.

One-sided tests: only values above OR below the hypothesized value p_0 is of interest. The alternative hypothesis becomes H_1: $p > p_0$ or H_1: $p < p_0$. z_0 becomes qnorm(1 - γ%). p value becomes pnorm(z, lower.tail=F) or pnorm(z). Others are the same.