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 $D_i = x_i - \bar{x}$.

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

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

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: $\bar{x} < \tilde{x}$, right skewed data: $\bar{x} > \tilde{x}$, symmetric data: $\bar{x} = \tilde{x}$.

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

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

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

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: $Z_i = {x_i - \bar{x}}/{SD}$.

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). $\tilde{x} = Q_2$. $IQR = Q_3 - Q_1$. 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 $N(\mu, \sigma^2)$.

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 $N(\mu, \sigma^2)$.

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

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).

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

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 $(\bar{x} + SD_x,\ \bar{y} + r \cdot SD_y)$.

Regression 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:

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{n \cdot E(X)}{n} = E(X)$$

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

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:

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

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

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:

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

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

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

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

For 0-1 box, $E(\bar{X}) = \mu = p,\ SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}.$ (p here is the proportion of 1s in the box, not the p value)

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.

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

Topic 09 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: $p = P(Z < -|z|) + P(Z > |z|) = P(2 \cdot Z > |z|) = 2 * pnorm(abs(z), lower.tail = F)$.

Z-test for 0-1 box:

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:

$|z| = \frac{|\bar{X}-p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} \le z_0$, 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.

Critical regions: the interval where we reject H_0. Critical region = (-∞, ∞) – Confidence interval.

HATPC: Hypotheses, Assumptions, Test Statistic, P-value, Conclusion.

Topic 10 T-tests

When $SE_0(\bar{x})$ is **unknown**, we estimate it using sample SD, which is called T-test.

$$T = \frac{\bar{X}-\mu_0}{\widehat{SE}_0(\bar{X})} = \frac{\bar{X}-\mu_0}{\frac{\hat{\sigma}}{\sqrt{n}}}$$

where

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}$$

The distribution curve of T test is similar to Z test (standard normal distribution curve), but the tails are fatter. The bigger degrees of freedom, the more similar.

Student's t-distribution:

- The "density" is computed using dt(x, df = n-1).
- Tail areas are computed using pt(x, df = n-1).
- Quantiles may be obtained using e.g. qt(0.975, df = m-1)

The confidence interval is given by:

$$\bar{x} \pm q\frac{\hat{\sigma}}{\sqrt{n}}$$

where q is the appropriate multiplier obtained using qt(), e.g. qt(0.975, df = 99).

If the box is not "nearly normal", we can try to approximate the distribution of T by simulating from a box which is "reasonably close" to the "real" box. This is known as the bootstrap principle.

We can also use the bootstrap principle to construct confidence intervals via simulation. $P\{l \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le u\} \approx 0.95$ and we may not have $l = -u$.

$(X-\mu)^2$ is a random draw from a different box. $E((X-\mu)^2) = \sigma^2 = SE(X)^2$

Topic 11 Two-Sample T-tests

We can model two sample groups from two separate boxes (independently of each other). First group: $X_1, \cdots, X_n$ taken (with repl.) from a box with mean μ_X and SD σ_X. Second group: $Y_1, \cdots, Y_n$ taken (with repl.) from a box with mean μ_Y and SD σ_Y.

$$E(\bar{X}-\bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_X - \mu_Y$$

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

Two-sample Test Statistics:

Null Hypothesis $H_0: \mu_X = \mu_Y$

If the σ_X and σ_Y were known:

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

If σ_X and σ_Y were unknown and $\sigma_X = \sigma_Y = \sigma$, Classical Two-Sample T-test.

If σ_X and σ_Y were unknown and not necessarily equal, Welch Test.

Classical Two-Sample T-test:

If $\sigma_X = \sigma_Y = \sigma$ and both boxes are approximately normal-shaped,

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

where

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

is called pooled estimate of σ (weighted average of $\hat{\sigma}_X^2$ and $\hat{\sigma}_Y^2$).

The confidence interval is (e.g. confidence level 95%):

$$P\left(l \le \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{SE(\bar{X}-\bar{Y})} \le u\right) = 0.95$$

$$P\big((\bar{X}-\bar{Y}) - u \times SE(\bar{X}-\bar{Y}) \le \mu_X - \mu_Y \le (\bar{X}-\bar{Y}) - l \times SE(\bar{X}-\bar{Y})\big) = 0.95$$

$$l = -u, u = qt(1-\frac{0.95}{2}, df = m+n-2)$$

Welch Test:

It assumes the two boxes are "approximately normal".

It uses Student's t-test whose degrees of freedom is a complicated function of m, n, σ_X and σ_Y.

Welch Test Using Stimulation:

If the two boxes are not "approximately normal", we can simulate from two surrogate boxes with equal means.

The p-value is:

$$P(\textit{simulation-based Welch statistic} \ge \textit{original Welch statistic})$$

```
mean(abs(Welch.stats.sim) >= abs(stat))
```

```
[1] 0.0936
```

We use simulated values to approximate the "true distribution" of the Welch statistic. So, confidence interval is given by "quantile" function then $E(\bar{X}-\bar{Y}) - interval \times SE(\bar{X}-\bar{Y})$

```
u.l = quantile(Welch.stats.sim, prob=c(.975, .025))
u.l
```

```
   97.5%      2.5%
2.017352 -2.160676
```

- That these are not the same magnitude indicates the slight lack of symmetry.

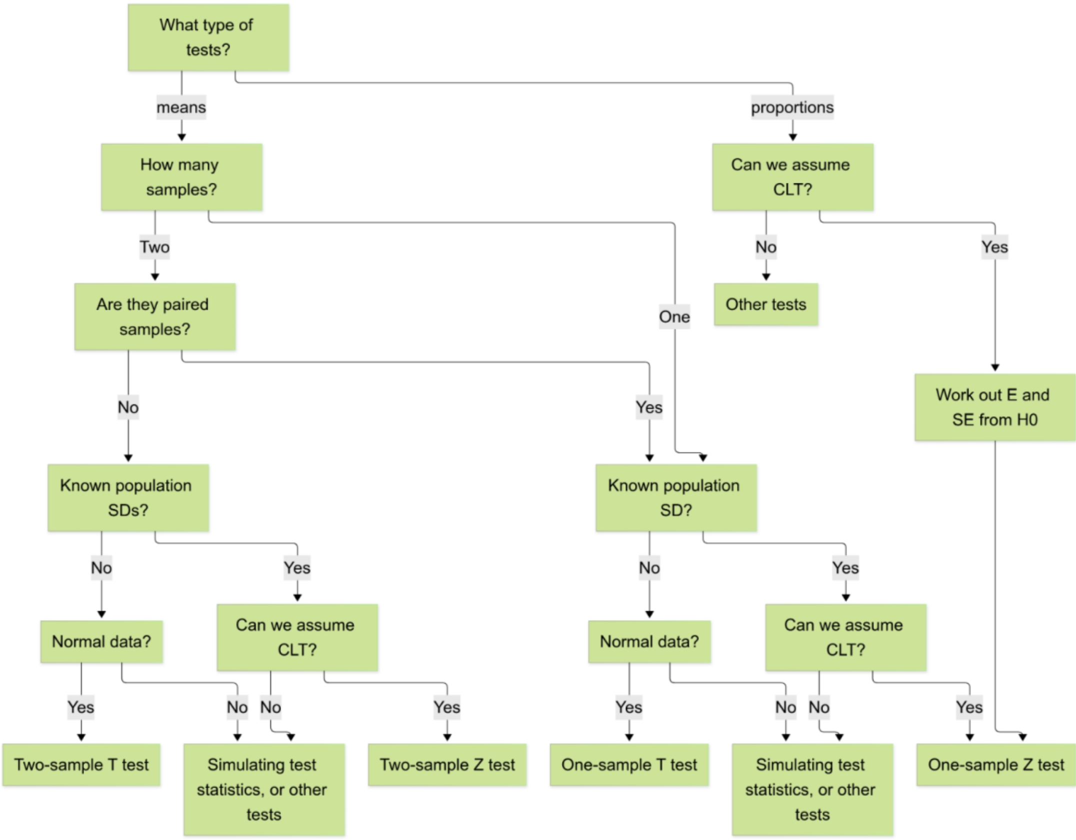
```
mean.diff - u.l*est.SE
```

```
     97.5%       2.5%
-14.303296   1.537169
```

Paired (two-sample) T-test:

Two samples of data (X, Y) are obtained from reading a pair of data (X_i, Y_i) from n individuals (not independent).

Null hypothesis $H_0: \mu_X = \mu_Y$, then perform T-test on the sample differences.



- One-sample Z test

$$Z = \frac{\bar{x} - E_0(\bar{X})}{SE_0(\bar{X})} \quad \text{where} \quad SE_0(\bar{X}) = \underbrace{\frac{\sigma}{\sqrt{n}}}_{\text{mean, known popu. SD}} \quad \text{or} \quad SE_0(\bar{X}) = \underbrace{\sqrt{\frac{p_0(1-p_0)}{n}}}_{\text{proportion}}$$

- One sample T test

$$T = \frac{\bar{x} - E_0(\bar{X})}{\frac{\hat{\sigma}}{\sqrt{n}}} \sim t_{n-1}$$

- Two-sample Z test

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

- Two-sample T test (classic)

$$T = \frac{\bar{X} - \bar{Y}}{\hat{\sigma}_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \sim t_{m+n-2}, \quad \hat{\sigma}_p = \sqrt{\frac{(m-1)\hat{\sigma}_X^2 + (n-1)\hat{\sigma}_Y^2}{m+n-2}}$$

- Two-sample T test (Welch)

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\hat{\sigma}_X^2}{m} + \frac{\hat{\sigma}_Y^2}{n}}} \sim t_{\text{dof}}$$

where the degrees of freesom (`dof`) is a complicated function of sample sizes and SDs (so we use R).

Topic 12 Chi-squared tests

Suppose we have data $X_1, \cdots, X_n$ only taking k distinct values (categories), modelled as a random sample taken with replacement from a box. We use integers $j = 1,2, \cdots k$ to label the categories.

$p_j = P(x = j)$ is the proportion of tickets in box labelled j.

$$\boldsymbol{p} = (p_1, \cdots p_k)$$

Null Hypothesis $H_0: \boldsymbol{p} = \boldsymbol{p_0}$ for hypothesized $\boldsymbol{P_0} = (p_{01}, \cdots p_{0k})$.

Alternative Hypothesis $H_1: not\ H_0$.

Observed frequencies are O_j, the number of data points labelled j.

The expected frequencies under H_0 are $E_j = np_{0j}$.

Pearson’s chi-squared statistic:

$$T = \sum_{j=1}^{k} \frac{\left(O_j - E_j\right)^2}{E_j}$$

Under H_0, for large n:

$$T \overset{approx.}{\sim} \chi_{k-1}^2$$

The chi-squared distribution with $k-1$ degrees of freedom.

Suppose the observed value of Pearson’s statistic is t_{obs}. The larger t_{obs}, the more evidence against H_0.

P-value is s given by the area under the χ_{k-1}^2 curve to the right of t_{obs}.

pchisq(…, df = …, lower.tail = F)

The χ_d^2 distribution:

Suppose we take d independent (i.e. with replacement) random draws from a N(0, 1) box: $Z_1, \cdots Z_d$. Then, $\sum_{i=1}^{d} Z_i^2$ has a χ_d^2 distribution. It is a skewed (to the right) distribution, but gets more symmetric as d increases.

The chi-squared test require: the sample size n is large, and expected frequencies are all at least 5.

For 0-1 box, chi-squared test is equivalent to two-sided Z-test for proportion.

Pearson's chi-squared using simulation:

```
sim.stat=0 # the dice example
for(i in 1:100000) {
   sim.rolls=sample(1:6, size=60, replace=T)
   freqs = tabulate(sim.rolls, nbins=6)  # works even with zero freqs, better than table()
   sim.stat[i] = chisq.test(freqs)$stat  # save the test statistics
}
```

- The observed Pearson statistic

```
Oi = table(die)
Ei = rep(10, 6)
rbind( Ei, Oi)
```

```
    1  2  3  4  5  6
Ei 10 10 10 10 10 10
Oi  4  6 17 16  8  9
```

```
stat=sum(((Oi-Ei)^2)/Ei)
stat
```

```
[1] 14.2
```

- P-value obtained using the simulated test distribution
  - Note that it's a one-sided test.

```
mean(sim.stat>=stat)
```

```
[1] 0.0139
```

Two-way tables: test of independence

Null Hypothesis: the events {being in Row i} and {being in Col j} are independent. That is $H_0: p_{ij} = P\{in\ Row\ i\ and\ Col\ j\} = P\{in\ Row\ i\} \times P\{in\ Col\ j\} = p_{i\cdot}p_{\cdot j}$.

Expected probability under H_0:

	Col 1	Col 2	$\cdots$	Col c	Total
Row 1	$p_{1\bullet}p_{\bullet 1}$	$p_{1\bullet}p_{\bullet 2}$	$\cdots$	$p_{1\bullet}p_{\bullet c}$	$p_{1\bullet}$
Row 2	$p_{2\bullet}p_{\bullet 1}$	$p_{2\bullet}p_{\bullet 2}$	$\cdots$	$p_{2\bullet}p_{\bullet c}$	$p_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	$p_{r\bullet}p_{\bullet 1}$	$p_{r\bullet}p_{\bullet 2}$	$\cdots$	$p_{r\bullet}p_{\bullet c}$	$p_{r\bullet}$
Total	$p_{\bullet 1}$	$p_{\bullet 2}$	$\cdots$	$p_{\bullet c}$	1

Expected frequencies under H_0 $(E_{ij} = np_{i.}p_{.j})$:

	Col 1	Col 2	$\cdots$	Col c	Total
Row 1	$np_{1\bullet}p_{\bullet 1}$	$np_{1\bullet}p_{\bullet 2}$	$\cdots$	$np_{1\bullet}p_{\bullet c}$	$np_{1\bullet}$
Row 2	$np_{2\bullet}p_{\bullet 1}$	$np_{2\bullet}p_{\bullet 2}$	$\cdots$	$np_{2\bullet}p_{\bullet c}$	$np_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	$np_{r\bullet}p_{\bullet 1}$	$np_{r\bullet}p_{\bullet 2}$	$\cdots$	$np_{r\bullet}p_{\bullet c}$	$np_{r\bullet}$
Total	$np_{\bullet 1}$	$np_{\bullet 2}$	$\cdots$	$np_{\bullet c}$	n

Observed frequencies:

	Col 1	Col 2	$\cdots$	Col c	Total
Row 1	O_{11}	O_{12}	$\cdots$	O_{1c}	$O_{1\bullet}$
Row 2	O_{21}	O_{22}	$\cdots$	O_{2c}	$O_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Row r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	$O_{r\bullet}$
Total	$O_{\bullet 1}$	$O_{\bullet 2}$	$\cdots$	$O_{\bullet c}$	n

Then, use chi-squared test

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

with degrees of freedom being

$$(r-1)(c-1)$$

Topic 13 Multiple linear regression

The simple linear regression model aims to predict the outcome of a dependent / response variable, which is a random draw Y, using an independent / explanatory variable x_1 and the model: $Y_i = b_0 + b_1x_{1i} + \varepsilon_i$.

The error ε_i are random draws taken from an "error box" with mean 0 and a fixed SE σ. The regression line $b_0 + b_1x_{1i}$ is the expected value of Y_i.

Simple linear regression with t-test:

Model: $Y_i = b_0 + b_1 x_{1i} + \varepsilon_i$

Null Hypothesis $H_0: b_1 = 0$ there is no linear relationship between x_1 and Y.

Assumptions: ε_i are independently drawn from the error box. $\varepsilon_i \sim (iid)\, N(0,\ \sigma^2)$ (iid stands for independent and identically distributed) (σ is the SD of the error box)

T-statistic:

$$T = \frac{\hat{b}_1 - b_1}{\widehat{SE}(\hat{b}_1)} = \frac{\hat{b}_1}{\widehat{SE}(\hat{b}_1)} \sim t_{n-2}$$

where

$$\widehat{SE}(\hat{b}_j) = \frac{\hat{\sigma}}{\sqrt{SST\ in\ x_1}} = \sqrt{\frac{1}{n-(p+1)} \frac{SSE}{SST\ in\ x_1}}$$

$$= \sqrt{\frac{1}{n-(p+1)} \frac{\sum_{i=1}^{n}\left(y_i - (\hat{b}_0 + \hat{b}_1 x_{1i})\right)^2}{\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2}}$$

Confidence intervals for regression coefficients (e.g. confidence level 99%):

Since

$$P\left(l \le \frac{\hat{b}_1 - b_1}{\widehat{SE}(\hat{b}_1)} \le u\right) = 0.99$$

$$l = -u$$

We have

$$P\left(\hat{b}_1 - u \times \widehat{SE}(\hat{b}_1) \le b_1 \le \hat{b}_1 + u \times \widehat{SE}(\hat{b}_1)\right) = 0.99$$

u is calculated by $u = qt(0.995,\ df = n-2)$

If we have multiple independent variables $x_1, x_2, \cdots, x_p$, the linear model becomes

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + \cdots + \hat{b}_p x_p$$

$$Y_i = \hat{b}_0 + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \cdots + \hat{b}_p x_{pi} + \varepsilon_i$$

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \varepsilon$$

where

$$\boldsymbol{Y} = (Y_1, Y_2, \cdots, Y_n)'$$

$$\boldsymbol{\beta} = (b_0, b_1, \cdots, b_p)'$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x_1}' \\ \boldsymbol{x_2}' \\ \vdots \\ \boldsymbol{x_n}' \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & \vdots & \vdots & \square & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix}$$

$$\varepsilon \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$$

Transformation:

If we see a non-linear relationship between y and x, we might be able to transform the data so that we have a linear relationship between the transformed variable(s). e.g. log(y) and x might have a linear relationship.

The optimal fit of multiple linear regression is

$$\begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_p \end{bmatrix} = \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

The coefficient of determination of multiple linear regression is

$$r^2 = 1 - \frac{SSE}{SST} = cor(y, \hat{y})^2$$

Multiple linear regression with t-test:

Null Hypothesis $H_0: b_j = 0$ $(j \in \{1,2, \cdots p\})$ there is no linear relationship between x_j and Y, after adjusting for all other independent variables in the model.

Equivalently, there is no linear relationship between x_j and U.

$$U_i = Y_i - (b_0 + b_1 x_{1i} + \cdots + b_{j-1} x_{j-1i} + b_{j+1} x_{j+1i} + \cdots + b_p x_p)$$

$$U_i = b_j x_{ji} + \varepsilon_i$$

Assumptions: ε_i are independently drawn from the error box. $\varepsilon_i \sim (iid)\, N(0,\ \sigma^2)$ (iid stands for independent and identically distributed) (σ is the SD of the error box)

T-statistic:

$$T = \frac{\hat{b}_j - b_j}{\widehat{SE}(\hat{b}_j)} = \frac{\hat{b}_j}{\widehat{SE}(\hat{b}_j)} \sim t_{n-(p+1)}$$

where

$$\widehat{SE}(\hat{b}_j) = \hat{\sigma} \times \sqrt{[(\boldsymbol{X}'\boldsymbol{X})^{-1}]_{jj}}$$

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-(p+1)}}$$

Topic 14 Model selection and logistic regression

The F-test is used to assess two nested models, where the null model is a special case of a more complicated alternative model containing additional independent variables.

→ From example, some of the possible models for the air pollution data are

$$\begin{aligned}
&\text{Model 4:} \quad \log(\text{ozone}_i) = b_0 + b_1 \cdot \text{radiation}_i + b_2 \cdot \text{temperature}_i + b_3 \cdot \text{wind}_i + \varepsilon_i \\
&\text{Model 3:} \quad \log(\text{ozone}_i) = b_0 + b_1 \cdot \text{radiation}_i + b_2 \cdot \text{temperature}_i + \varepsilon_i \\
&\text{Model 2:} \quad \log(\text{ozone}_i) = b_0 + b_2 \cdot \text{temperature}_i + \varepsilon_i \\
&\text{Model 1:} \quad \log(\text{ozone}_i) = b_0 + \varepsilon_i
\end{aligned}$$

- When Model 1 is the null model, Model 2, 3, or 4 can be a valid alternative model
- When Model 1 is the null model, Model 3 or 4 can be a valid alternative model

We may want to test whether the additional independent variables in the alternative model significantly improve the fit of the null model.

F-test Assumptions: same as multiple linear regression.

Overall F-test:

Null hypothesis $H_0: b_1 = b_2 = \cdots = b_p = 0$ all regression coefficients (except the intercept) are zero. That is: $Y_i = b_0 + \varepsilon_i$

Alternative hypothesis H_1: at least one of the regression coefficients is not zero.

Partial F-test:

Null hypothesis $H_0: b_1 = b_2 = 0$ some regression coefficients (except the intercept) are zero. The additional independent variables (x_1 and x_2 in this example) have no effect in explaining Y. That is: $Y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \varepsilon_i$

Alternative hypothesis H_1: at least one of the additional independent variables has an effect in explaining Y.

Consider a null model with q independent variables and an alternative model with p

independent variables. The alternative model is always larger, so $p > q$.

Under H_0: Fit the model and calculate $\widehat{SSE}_{H_0}$. Degrees of freedom is $n-(q+1)$

Under H_1: Fit the model and calculate $\widehat{SSE}_{H_1}$. Degrees of freedom is $n-(p+1)$

The F statistic:

$$F = \frac{(\widehat{SSE}_{H_0} - \widehat{SSE}_{H_1})/(p-q)}{\widehat{SSE}_{H_1}/(n-(p+1))} \sim F_{p-q,n-(p+1)}$$

Numerator: explained variation per additional independent variable.

Denominator: unexplained variation in the alternative model per degree of freedom

One-sided test, only large values of F argue against H_0.

Adjusted R-squared:

$$\begin{aligned}
&Adjusted\ R\text{-}squared \\
&= 1 - \frac{Estimated\ SD\ of\ the\ residual\ error}{Sample\ SD\ of\ the\ \ dependent\ variable} \\
&= 1 - \frac{\hat{\sigma}}{\hat{s}_X} \\
&= 1 - \frac{\widehat{SSE}/(n-(p+1))}{\widehat{SST}/(n-1)} \\
&= 1 - (1-r^2)\frac{n-1}{n-(p+1)} \\
&\geq r^2
\end{aligned}$$

Adjusted R-squared penalizes the inclusion of unhelpful independent variables.

In choosing between models, statisticians have two aims:

To choose a simple (i.e. not too complex) model. A possibility to measure the complexity of a linear regression model is by the number of independent variables, p. The greater this value, the more complex the model.

To choose a model that fits the data well. Possibilities to measure the closeness of fit of the model to data are R-squared, adjusted R-squared, etc.

Backward variable selection

We start with a full model containing all possible independent variables. In each iteration of the backward variable selection:

1. Start with the current model, for each independent variable in turn, investigate the effect of removing a variable from the current model.
2. Remove the least significant variable, unless this independent variable is supplying significant information about the dependent variable Y.
3. Go to step 1. Stop only if all variables in the current model are important.

Forward variable selection

We start with the model containing no independent variables, i.e., the baseline model $\hat{y} = \bar{y}$. In each iteration of the forward variable selection:

1. For each variable in turn, investigate the effect of adding an independent variable to the current model.
2. Add the most informative variable, unless this variable is not supplying significant information about the dependent variable Y.
3. Go to step 1. Stop only if all of the non-included variables are not significant.

If an event is occurring with probability p, its odds is defined as:

$$odds = \frac{probability\ that\ event\ will\ occur}{probability\ that\ event\ will\ not\ occur} = \frac{p}{1-p}$$

Its logit is defined as:

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

Binomial Distribution:

Bernoulli trials: A probability experiment with only two possible outcomes and each trial has an independent and identical chance of success.

Binomial Distribution describes the probability of obtaining a certain number of successes in a fixed number of Bernoulli trials.

$$Y_i \sim Binomial(m_i, p_i)$$

where m is the number of trails and p is success chance.

Logistic Regression:

Use independent variables $x_1, \cdots x_p$ to **predict logit** instead of directly predicting success or not.

$$logit(p_i) = \log \frac{p_i}{1 - p_i} = b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_p x_{pi}$$

which also gives

$$Y_i \sim Binomial(m_i, \frac{odds_i}{1 + odds_i})$$

where

$$odds_i = \mathrm{n}^{logit(p_i)}$$

Deviance is used to measure the quality of the model fit, lower deviance means better fit.