Formulae: Statistical Computing

Master in Data Science, Winter Semester 2019/20 — Prof. Tim Downie
Last edit: January 5, 2020

1 Data Types

- Descriptive Statistics
  - Qualitative variables
    - Nominal
    - Ordinal
  - Numeric or Quantitative Variables
    - Discrete
    - Continuous
- Object types in R
  - Factor (Qualitative)
  - Numeric (Quantitative)
  - Logical
  - Character
  - List

2 Frequency

- Absolute Frequency h_i (`table()`)
- Relative Frequency $f_i = \dfrac{h_i}{n}$ (`prop.table(table())`)
- Absolute cumulative frequency $H_i = \sum_{j=1}^{i} h_j$ (`cumsum(table())`)
- Relative cumulative frequency $F_i = \sum_{j=1}^{i} f_j = \dfrac{H_i}{n}$ (`cumsum(prop.table(table()))`)

3 Descriptive Statictics

- Mean (arithmetic mean) $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
  - If $y_i = ax_i + b$ (a & b constant), then $\overline{y} = a\overline{x} + b$.
  - If $z_i = x_i + y_i$, then $\overline{z} = \overline{x} + \overline{y}$.
- Median $x_{0.5}$
  The ordered data values are $x_{(1)}, \ldots, x_{(n)}$
  - odd n: $x_{0.5} = x_{\left(\frac{n+1}{2}\right)}$
  - even n: $x_{0.5} = \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)$
- Mode x_D is the most frequent value.
- Variance $s_x^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$
- Standard deviation (SD) $s_x = \sqrt{s_x^2}$
  If $y_i = ax_i + b$ (a & b constant), then
  - Var $s_y^2 = a^2 s_x^2$
  - SD $s_y = a s_x$
- Range $R = x_{\text{max}} - x_{\text{min}}$
- Interquartile range $IQR = Q_3 - Q_1$
- Coefficient of variation $CV = \dfrac{s}{\overline{x}}$
- First quartile (Q_1)
  - If n is divisible by 4
    $Q_1 = x_{0.25} = \frac{1}{2}\left(x_{\left(\frac{n}{4}\right)} + x_{\left(\frac{n}{4}+1\right)}\right)$
  - If n is not divisible by 4 $Q_1 = x_{0.25} = x_{\left(\lceil \frac{n}{4} \rceil\right)}$
    $\lceil \cdot \rceil$ means round *up*.
  - R: `quantile(x,0.25)`
- Third quartile (Q_3)
  - If n is divisible by 4
    $Q_3 = x_{0.75} = \frac{1}{2}\left(x_{\left(\frac{3n}{4}\right)} + x_{\left(\frac{3n}{4}+1\right)}\right)$
  - If n is not divisible by 4 $Q_3 = x_{0.75} = x_{\left(\lceil \frac{3n}{4} \rceil\right)}$
  - R: `quantile(x,0.75)`
- p-quantile
  - If pn ist an integer $x_p = \frac{1}{2}\left(x_{(pn)} + x_{(pn+1)}\right)$
  - If pn ist not an integer $x_p = x_{(\lceil pn \rceil)}$
  - R: `quantile(x,p)`
- Skewness (Symmetry): g_1
  - $g_1 \gg 0$ right-skewed, right-tailed, leaning to the left
  - $g_1 \ll 0$ left-skewed, left-tailed, leaning to the right
  - $g_1 \approx 0$ symmetric.

- Covariance $s_{xy} = \frac{1}{n-1} \sum (x_i - \overline{x})(y_i - \overline{y})$
- Correlation coefficient

  $$r_{x,y} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2 \sum (y_i - \overline{y})^2}} = \frac{s_{xy}}{s_x . s_y}$$

- Empirical distribution function

  $$F_n(b) = P(X \leqslant b) = \frac{\# x_i \leqslant b}{n}$$

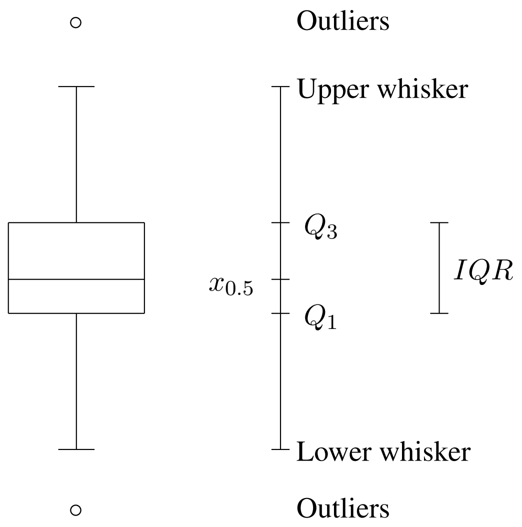
- R: `ecdf(x)`

4 Graphics

Histogram

Height of i-th Column is the "density" $y_i = \frac{h_i}{b_i \cdot n}$, where h_i is the absolute frequency in the i-th interval and b_i is the interval width. R: `hist(x)`

Box plot



Upper whisker is the largest data value $\leqslant Q_3 + 1.5IQR$

Lower whisker is the smallest data value $\geqslant Q_1 - 1.5IQR$

R: `boxplot(y)` or `boxplot(y~x)`

5 Normal Distribution

- Let $Z \sim N(0,1)$ be a random variable with the standard normal distribution, $P(Z \leqslant z) = \Phi(z)$
  R: `pnorm(z)`
- Let X have a general normal $N(\mu, \sigma^2)$ distribution. $Z = \frac{X - \mu}{\sigma}$ has a standard normal distribution.
- Central limit theorem: Let $X_1, X_2, \ldots, X_n$ be an iid. random sample, from an arbitrary distribution with expectation μ and variance σ^2

  For large n, the distribution of the random variable $Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2 / n}}$ is well approximated by the standard normal $N(0,1)$ distribution.

  $$\Rightarrow \qquad \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \overset{a}{\sim} N(0,1) \qquad \text{or equivalently} \qquad \overline{X} \overset{a}{\sim} N(\mu, \sigma^2/n)$$

6 Regression

Regression line for paired data (x_i, y_i):

$$y_i = \widehat{a} + \widehat{b}x_i + \widehat{\epsilon}_i,$$

where $\widehat{a}$ is the least squares estimator for the intercept and $\widehat{b}$ is the least squares estimate for the gradient. $\widehat{\epsilon}_i$ is the i-th residual or i-th error term.

The regression coefficients are calculated using:

- $\widehat{b} = \dfrac{\sum(y_i - \overline{y})(x_i - \overline{x})}{\sum(x_i - \overline{x})^2} = \dfrac{s_{xy}}{s_x^2}$
- $\widehat{a} = \overline{y} - \widehat{b}\overline{x}$

The fitted values are $\widehat{y}_i = \widehat{a} + \widehat{b}x_i$.
The residuals are $\widehat{\epsilon}_i = y_i - \widehat{y}_i$.

7 Confidence intervals

- A confidence interval for μ with 95% confidence level, based on the normal distribution

  $$\left[\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}}\right],$$

  estimate σ using s_x if σ is unknown. For other confidence levels $(1-\alpha)$ use `qnorm(1-alpha/2)`.

- A confidence interval for μ with 95% confidence level, based on the t distribution

  $$\left[\overline{x} \pm t\frac{s_x}{\sqrt{n}}\right]$$

  t depends on the confidence level and the sample size `qt(1-alpha/2,n-1)`.

- **Confidence interval for a proportion** p
  $\widehat{p} = \overline{x} = \dfrac{\sum_{i=1}^n x_i}{n}$ is an estimate for p.

  $$\left[\overline{x} \pm 1.96\frac{\sqrt{\overline{x}(1-\overline{x})}}{\sqrt{n}}\right]$$

  is an approximate 95% confidence interval for p, provided $n > 30$.

8 Hypothesis tests

One sample t-tests `t.test(x)`

- Two sided test for an expectation μ with significance level α

  $$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

  Critical region: H_0 is rejected iff $t_{\text{stat}} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} > t_{cr}$ or $t_{\text{stat}} < -t_{cr}$,
  where t_{cr} is a quantile from the t-distribution `qt(1-alpha/2,n-1)`.

- One sided t-test for an expectation μ with significance level α

  a) $H_0 : \mu \geqslant \mu_0$ vs $H_1 : \mu < \mu_0$ Critical region: H_0 is rejected iff $t_{\text{stat}} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} < -t_{cr}$,
  t_{cr} is `qt(1-alpha,n-1)`

  b) $H_0 : \mu \leqslant \mu_0$ vs $H_1 : \mu > \mu_0$ H_0 is rejected iff $t_{\text{stat}} = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} > t_{cr}$ t_{cr} is `qt(1-alpha,n-1)`

- p-Value: reject the null hypothesis iff $p<\alpha$ the significance level.

Two sample tests `t.test(x,y) t.test(x~y)`

- For two unpaired samples:
  Test statistic: $t_{\text{stat}} = \dfrac{\overline{x} - \overline{y}}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}$, with pooled variance $s_p^2 = \dfrac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}$.
  Degrees of freedom: $m = n_x + n_y - 2$.
- For two unpaired samples: Calculate $d_i = x_i - y_i$ and carry out a one sample t-test on d_i.
- Critical region for two sided tests $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x \neq \mu_y$
  H_0 is rejected iff $t_{\text{stat}} > t_{cr}$ or $t_{\text{stat}} < -t_{cr}$.
- Critical region for one sided tests
  (a) $H_0 : \mu_x \geqslant \mu_y$ vs $H_1 : \mu_x < \mu_y$. $\Rightarrow H_0$is rejected iff $t_{\text{stat}} < -t_{cr}$.
  (b) $H_0 : \mu_x \leqslant \mu_y$ vs $H_1 : \mu_x > \mu_y$. $\Rightarrow H_0$ is rejected iff $t_{\text{stat}} > t_{cr}$.

χ^2 Test of independence

For variables X and Y with values $X_1, \ldots X_m$ and $Y_1, \ldots, Y_n$. Joint frequency table:

	Y_1	$\cdots$	Y_n	Total
X_1	h_{11}	$\cdots$	h_{1n}	$h_{1\cdot}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
X_m	h_{m1}	$\cdots$	h_{mn}	$h_{m\cdot}$
Total	$h_{\cdot 1}$	$\cdots$	$h_{\cdot n}$	n

The expected frequencies are: $e_{ij} = \dfrac{h_i.h._j}{n}$ The test statistic is: $\chi^2_{\text{stat}} = \sum_{i,j} \dfrac{(h_{ij} - e_{ij})^2}{e_{ij}}$
Degrees of freedom: $k = (m_X - 1)(m_Y - 1)$
Critical value is the $1 - \alpha$-quantile from the χ^2_k distribution `qchisq(1-alpha,k)`.
H_0 is rejected iff $\chi^2_{\text{stat}} >$ critical value.

Test of equality of two variances

The test statistic is $f_{stat} = \dfrac{s_x^2}{s_y^2}$ R: `var.test(x,y)` or `var.test(x y)`