1 Descriptive Statistics

Types of Variables

- Categorical/Qualitative: Nomial, Ordinal
- Quantitative: Discrete, Continuous

Sample Median:

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & x \text{ is odd} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n+1}{2}}) & x \text{ is even} \end{cases}$$

Trimmed Mean: Eliminate $x\%$ smallest and largest variables and take the mean.

Sample Variance:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

p**th Quartile:** Let $m = np + 0.5$. Take the average of the $\lfloor m \rfloor$th and $\lceil m \rceil$th smallest data points.

Outlier: $\leq Q1 - 1.5IQR$ or $\geq Q3 + 1.5IQR$

2 Random Variables

Chebyshev's Theorem:

$$\begin{aligned} P(|X-\mu| \geq t) &= \int_{|x-\mu|\geq t} f(x)\mathrm{d}x \\ &\leq \int_{|x-\mu|\geq t} \frac{(x-t)^2}{t^2} f(x)\mathrm{d}x \leq \frac{\sigma^2}{t^2} \end{aligned}$$

Geometric: Expectation $\frac{1}{p}$, variance $\frac{1-p}{p^2}$, and

$$p(k) = p \cdot (1-p)^{k-1}$$

Binomial: Expectation np, variance $np(1-p)$, and

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Poisson: If $\lambda > 0$ is the occurences of an event per unit time, then $X \sim \text{Poisson}(\lambda)$, expectation and variance λ, p.m.f.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \text{ for } k \geq 0$$

Theorem. The average occurences Y_t over t units of time, with rate λ per unit time:

$$Y_t \sim \text{Poisson}(\lambda t)$$

Poisson Limit Theorem. Suppose $p_n \in [0,1]$ and np_n converges to λ. Then

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{1-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

Normal: For μ, σ^2, if $X \sim N(\mu, \sigma^2)$, we have the pdf:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$X \sim N(\mu, \sigma^2) \implies \frac{X-\mu}{\sigma} = Z \sim N(0,1)$$

3 Moments

The rth population moment is $E(X^r)$. The rth central moment is $\mu_r = E([X - E(X)]^r)$.

Skewness: $E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ *Symmetry of Distribution*

(Negative) Left-skewed: Data concentrated to the right. Mean lower than median lower than mode

Kurtosis: $E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$ *Tail Behaviour*

Compare with a normal distribution with kurtosis 3. Excess kurtosis is Kurtosis - 3. Positive excess means thicker tails.

MGF:

$$M_X(t) = E(e^{tX}) = \begin{cases} \displaystyle\sum_x e^{tx} p_X(x) & X \text{ is discrete} \\ \displaystyle\int_{-\infty}^{+\infty} e^{tx} f_X(x)\mathrm{d}x & X \text{ is continuous} \end{cases}$$

Use to find population moments:

$$M_X^k(0) = \frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t)\Big|_{t=0} = E(X^k)$$

Binomial: $M_X(t) = (pe^t + 1 - p)^n$

Normal: $M_Z(t) = e^{-\frac{1}{2}t^2}$. For transformations:

$$M_{aX+b}(t) = E(e^{t(aX+b)}) = e^{t\mu} E(e^{(tb)Z}) = e^{tu} M_X(bt)$$

4 Sampling Distribution Theorems

For i.i.d $X_i \sim N(0,1)$, $Z \sim N(0,1)$, $Y \sim \chi^2(n)$, all independent:

$$\sum_{i=1}^{n} X_i^2 \sim \chi^2(n) \qquad W = \frac{Z}{\sqrt{Y/n}} \sim t(n)$$

For i.i.d $X_i \sim N(\mu_X, \sigma_X^2)$:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right) \qquad \frac{(n-1)S_{n-1}^2}{\sigma_X^2} \sim \chi^2(n-1)$$

and $S_{n-1}^2, \bar{X}$ are independent. Also:

$$\frac{\bar{X} - \mu_X}{S_{n-1}/\sqrt{n}} \sim t(n-1)$$

Inverses of distributions where $Z \sim N(0,1)$, $T_n \sim t(n)$, $U_n \sim \chi_n^2$:

$$z_\alpha := P(Z > z_\alpha) = \alpha \qquad t_{n,\alpha} := P(T_n > t_{n,\alpha}) = \alpha$$
$$\chi_{n,\alpha}^2 := P(U_{n-1} > \chi_{n,\alpha}^2) = \alpha$$
$$P(Z > a) = P(Z < -a) = 1 - P(Z < a)$$

The t-distribution is also symmetric.

5 Parameter Estimation

$$E(\overline{X^k}) = E(X^k) \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \qquad E(S_{n-1}) = \sigma^2$$

Interval μ_X with known σ_X^2:

$$\left[\bar{X} - z_{\frac{\alpha}{2}}\frac{\sigma_X}{\sqrt{n}}, \bar{X} + z_{\frac{\alpha}{2}}\frac{\sigma_X}{\sqrt{n}},\right]$$

Interval μ_X with unknown σ_X^2, $P(T_{n-1} > t_{n-1,\alpha/2}) = \alpha$:

$$\left[\bar{X} - t_{n-1,\alpha/2}\frac{S_{n-1}}{\sqrt{n}}, \bar{X} + t_{n-1,\alpha/2}\frac{S_{n-1}}{\sqrt{n}}\right]$$

Interval σ_X^2 with unknown μ_X (**take square root for σ_X**):

$$\left[\frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,\alpha/2}}, \frac{(n-1)S_{n-1}^2}{\chi^2_{n-1,1-\alpha/2}}\right]$$

NOTE: χ^2 distribution is not symmetric.

Interval σ_X^2 with known μ_X (note that we have n DOF instead of $n-1$):

$$\left[\frac{\sum_{i=1}^n (X_i-\mu_X)^2}{\chi^2_{n,\alpha/2}}, \frac{\sum_{i=1}^n (X_i-\mu_X)^2}{\chi^2_{n,1-\alpha/2}}\right]$$

6 Hypothesis Testing

	Not reject H_0	Reject H_0
H_0 *true*	No error	**Type I error**
H_0 *false*	**Type II error**	No error

Power of test statement: $1-\beta$

Tests for μ_X: (for p-value, reject if less than α)

RY	$\bar{x} > \mu_0 + z_\alpha \frac{\sigma_X}{\sqrt{n}}$	$P\left(Z > \frac{\bar{x}-\mu_0}{\sigma_X/\sqrt{n}}\right)$
LY	$\bar{x} < \mu_0 - z_\alpha \frac{\sigma_X}{\sqrt{n}}$	$P\left(Z < \frac{\bar{x}-\mu_0}{\sigma_X/\sqrt{n}}\right)$
TY	$\left\lvert\frac{\bar{x}-\mu_0}{\sigma_X/\sqrt{n}}\right\rvert > z_{\alpha/2}$	$2P\left(Z > \left\lvert\frac{\bar{x}-\mu_0}{\sigma_X/\sqrt{n}}\right\rvert\right)$
RN	$\bar{x} > \mu_0 + t_{n-1,\alpha} \frac{s_{n-1}}{\sqrt{n}}$	$P\left(T_{n-1} > \frac{\bar{x}-\mu_0}{s_{n-1}/\sqrt{n}}\right)$
LN	$\bar{x} < \mu_0 - t_{n-1,\alpha} \frac{s_{n-1}}{\sqrt{n}}$	$P\left(T_{n-1} < \frac{\bar{x}-\mu_0}{s_{n-1}/\sqrt{n}}\right)$
TN	$\left\lvert\frac{\bar{x}-\mu_0}{s_{n-1}/\sqrt{n}}\right\rvert > t_{n-1,\alpha/2}$	$2P\left(T_{n-1} > \left\lvert\frac{\bar{x}-\mu_0}{s_{n-1}/\sqrt{n}}\right\rvert\right)$

Tests for σ_X^2 (μ_X unknown): let $U_n \sim \chi^2(n)$, then

RN	$\frac{(n-1)s_{n-1}^2}{\sigma_0^2} > \chi^2_{n-1,\alpha}$	$P\left(U_{n-1} > \frac{(n-1)s_{n-1}^2}{\sigma_0^2}\right)$
LN	$\frac{(n-1)s_{n-1}^2}{\sigma_0^2} < \chi^2_{n-1,1-\alpha}$	$P\left(U_{n-1} < \frac{(n-1)s_{n-1}^2}{\sigma_0^2}\right)$

Two-sided critical value:

$$\frac{(n-1)s_{n-1}^2}{\sigma_0^2} > \chi^2_{n-1,\alpha/2} \quad \text{or} \quad \frac{(n-1)s_{n-1}^2}{\sigma_0^2} < \chi^2_{n-1,1-\alpha/2}$$

p-value:

$$2\min\left\{P\left(U_{n-1} > \frac{(n-1)s_{n-1}^2}{\sigma_0^2}\right), P\left(U_{n-1} < \frac{(n-1)s_{n-1}^2}{\sigma_0^2}\right)\right\}$$

If μ_X is known, replace $(n-1)S_{n-1}^2$ with $\sum_{i=1}^n (x_i-\mu_X)^2$.

7 Simple Linear Regression

Model: $Y = \beta_0 + \beta_1 x + \varepsilon$ where ε is a random variable.

Denote:

$$S_{XY} = \sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{x}) = \sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}$$

Least Square Estimators:

$$b = \hat{\beta}_1 = \frac{S_{XY}}{S_{XX}} \qquad a = \hat{\beta}_0 = \bar{y} - b\bar{x}$$

Denote $\hat{y} = a + bx$, residual $e_i = y_i - \hat{y}_i$, $SSE = \sum_{i=1}^n e_i^2$.

Correlation Coefficient $r \in [-1,1]$:

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

which estimates

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

Confidence Interval: $z_{Fisher} = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$.

Random variable Z_{Fisher} approximately follows normal distribution with mean $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ and variance $\frac{1}{n-3}$. Construct CI for $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$.

Variability of Response: Partition: (RSS = Regression Sum of Squares, SST = Total Sum of Squares)

$$\underbrace{\sum_{i=1}^n (y_i-\bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^n (\hat{y}-\bar{y})^2}_{RSS} + \underbrace{\sum_{i=1}^n (y_i-\hat{y})^2}_{SSE}$$

SSE: Variability unexplained by the model. RSS: Variability explained by the model.

Proportion of Variability Explained: $R^2 = \frac{RSS}{SST} = 1 - \frac{SSE}{SST}$

For $Y = \beta_0 + \beta_1 x + \varepsilon$, $\varepsilon_i \sim N(0,\sigma^2)$ i.i.d, and $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$:

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{XX}}\right) \qquad \hat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{XX}}\right)$$

Can be used to construct CI and test hypotheses.

Estimate σ^2:

$$S^2 = \frac{\sum_{i=1}^n (Y_i-\hat{Y}_i)^2}{n-2} \qquad s^2 = \frac{\sum_{i=1}^n (y_i-\hat{y}_i)^2}{n-2} = \frac{S_{YY}-bS_{XY}}{n-2}$$

are unbiased estimators and estimations of σ^2.

Because

$$T_{n-2} = \frac{\hat{\beta}_1-\beta_1}{\sqrt{\frac{S^2}{S_{XX}}}} \sim t(n-2) \qquad T_{n-2} = \frac{\hat{\beta}_0-\beta_0}{\sqrt{\frac{S^2\sum_{i=1}^n x_i^2}{nS_{XX}}}} \sim t(n-2)$$

we have confidence intervals for β_1, β_0:

$$b \pm t_{n-2,\alpha/2}\sqrt{\frac{s^2}{S_{XX}}} \qquad a \pm t_{n-2,\alpha/2}\sqrt{\frac{s^2\sum_{i=1}^n x_i^2}{nS_{XX}}}$$

and for $H_0 = b_1$ or $H_0 = a_1$, t-values

$$t_b = \frac{b-b_1}{\sqrt{\frac{s^2}{S_{XX}}}} \qquad t_a = \frac{a-a_1}{\sqrt{\frac{s^2\sum_{i=1}^n x_i^2}{nS_{XX}}}}.$$

R	$t_x > t_{n-2,\alpha}$	$P(T_{n-2} > t_x) < \alpha$
L	$t_x < -t_{n-2,\alpha}$	$P(T_{n-2} < t_x) < \alpha$
T	$\lvert t_x\rvert > t_{n-2,\alpha/2}$	$2P(T_{n-2} > \lvert t_x\rvert) < \alpha$

Prediction: Prediction interval for y_{new}:

$$\hat{y}_{new} \pm t_{n-2,\alpha/2}\cdot s\sqrt{1+\frac{1}{n}+\frac{(x_{new}-\bar{x})^2}{S_{XX}}}$$

where $\hat{y}_{new} = a + bx_{new}$ and $s = \sqrt{s^2}$.

Integral Cheatsheet

$$\int \tan x \mathrm{d}x = -\ln|\cos x| \qquad \int \cot x \mathrm{d}x = \ln|\sin x|$$

$$\int \frac{1}{\sqrt{1-x^2}}\mathrm{d}x = \arcsin x \qquad \int \frac{1}{1+x^2}\mathrm{d}x = \arctan x$$

$$\int_0^\infty x^n e^{-x}\mathrm{d}x = \Gamma(n+1) = n!$$

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} x^n e^{-\frac{1}{2}x^2}\mathrm{d}x = \begin{cases} 0 & n \text{ odd} \\ 1\cdot 3\cdots(n-1) & n \text{ even} \end{cases}$$