Descriptive Statistics Summary

One Variable

mean $\bar{x} = \dfrac{\Sigma x_i}{n}$

standard deviation $s = \sqrt{\dfrac{1}{n-1}\Sigma(x_i - \bar{x})^2}$ median
quartiles

variance $= s^2$

outliers $= 1.5 \times \text{IQR}$

Two variables

x = explanatory variable y = response variable

Regression line $y = a + bx$

$b = \dfrac{rs_y}{s_x}$ minimizes $D^2 = \Sigma(y_i - \hat{y}_i)^2 = \Sigma(y_i - (a + bx_i))^2$

a is vertical intercept

$a = \bar{y} - b\bar{x}$ b is slope

Correlation, r Coefficient of determination, r^2

$r \approx +1$

$r \approx -1$ $r^2 = \dfrac{\text{Variance in predicted } \hat{y} \text{ values}}{\text{Variance in observed } y \text{ values}} = \dfrac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2}$

$r \approx 0$

Residuals $y_i - \hat{y}$ all close to 0 if line fits well

Sample statistics vs. population parameters

$\bar{x}, s$ μ, σ

Random Variables

Discrete $\mu = \Sigma\, p_i\, x_i$ $\quad \sigma = \sqrt{\Sigma\, p_i\, (x_i - \mu)^2}$ example: Binomial, $B(n,p)$

Continuous $\mu = \int_{-\infty}^{\infty} x f(x) dx$ $\quad \sigma = \sqrt{\int_{-\infty}^{\infty} f(x)\,(x-\mu)^2\, dx}$ example: Normal, $N(\mu,\sigma)$

Inference Summary

Sampling Distribution (approximately normal)

Means $\bar{x}$

$\mu_{\bar{x}} = \mu, \;\; \sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$

Proportions $\hat{p}$

$\mu_{\hat{p}} = p, \;\; \sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

Confidence Interval: Means	Confidence Interval: Proportions

one sample

σ known: $\left(\bar{x} - z\dfrac{\sigma}{\sqrt{n}}, \bar{x} + z\dfrac{\sigma}{\sqrt{n}} \right)$

one sample

$\hat{p} = \dfrac{X}{n}$

one sample

σ not known: $\left(\bar{x} - t\dfrac{s}{\sqrt{n}}, \bar{x} + t\dfrac{s}{\sqrt{n}} \right)$

$\left(\hat{p} - z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}},\; \hat{p} + z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} \right)$

two samples

σ not known:

$\left(\bar{x}_1 - \bar{x}_2 - t\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}},\; \bar{x}_1 - \bar{x}_2 + t\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}} \right)$

two samples

$\hat{p}_1 = \dfrac{X_1}{n_1} \qquad \hat{p}_2 = \dfrac{X_2}{n_2}$

$SE_{\tilde{p}} = \sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$

(matched pairs, find CI for differences using one sample methods)

$(\hat{p}_1 - \hat{p}_2 - z\, SE_{\hat{p}},\; \hat{p}_1 - \hat{p}_2 + z\, SE_{\hat{p}})$

Hypothesis Tests: Means	Hypothesis Tests: Proportions

one sample σ known: H_0 : mean = μ

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$ standard normal

one sample σ not known: H_0 : mean = μ

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$ t distrib, $df = n - 1$

two samples σ not known: $H_0 : \mu_1 = \mu_2$

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n}}}$$

t distrib, df = min $(n_1 - 1,\ n_2 - 1)$

many samples largest $s < 2 \cdot$ smallest s

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_K$

ANOVA $F(k-1,\ N-k)$ $F = \dfrac{MSG}{MSE}$

k groups, N data altogether

one sample H_0 : prop = p_0

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$ standard normal

two samples $H_0 : p_1 = p_2$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad \hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$$

std norm pooled sample

many samples H_0 : Independence

Chi Square (all counts ≥ 5)

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$