Random Variables

- Generally, when $f(x)$ is the pmf for (discrete) r.v. X:

$$\mathrm{E}(X^j) = \sum_x x^j f(x)$$

- The variance is $\quad \mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$

Sample statistics

For iid sample $X_1, \ldots, X_n$ from population with mean μ, sd σ,

$Y = \sum_i X_i$: $\quad \mathrm{E}(Y) = n\mu, \quad \mathrm{SD}(Y) = \sigma\sqrt{n}$ $\quad$ (multiply)

$\overline{X} = Y/n$: $\quad \mathrm{E}(\overline{X}) = \mu, \quad \mathrm{SD}(\overline{X}) = \sigma/\sqrt{n}$ $\quad$ (divide)

- When the population is normal, or n large, then

$$Y \sim \mathsf{Norm}(n\mu, \sigma\sqrt{n}) \quad \text{and} \quad \overline{X} \sim \mathsf{Norm}(\mu, \sigma/\sqrt{n}).$$

- When the X_i are Bernoulli (i.e., $\mathsf{Binom}(1, p)$), then $Y = X_1 + X_2 + \cdots + X_n \sim \mathsf{Binom}(n, p)$. Moreover, when $np \geq 10$ and $n(1-p) \geq 10$, then Y is approx. normal:

$$Y \sim \mathsf{Norm}(np, \sqrt{np(1-p)}) \quad \text{and}$$

$$\hat{p} = \frac{Y}{n} \sim \mathsf{Norm}(p, \sqrt{p(1-p)/n})$$

Inference Procedures

- Level C Confidence Intervals (general):

(estimate) ± (critical value)(approx. std. error)

- 1-sample proportion:
  - CIs for p, $\quad \mathrm{SE} = \sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$
  - z-test (when $\hat{p}$ approx. normal)

test stat. ($\mathbf{H}_0$: $p = p_0$): $z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

- 1-sample t: test statistic when $\mathbf{H}_0$: $\mu = \mu_0$

$$t = \frac{\overline{x} - \mu_0}{\mathrm{SE}}, \quad \mathrm{SE} = \frac{s}{\sqrt{n}}, \quad df = n - 1$$

- 2-sample t: test statistic when $\mathbf{H}_0$: $\mu_1 - \mu_2 = 0$

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\mathrm{SE}}, \quad \mathrm{SE} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Use t-distribution with conservative estimate for degrees of freedom $df = \min(n_1, n_2) - 1$

- Chi-square test statistic:

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

contingency table: df = (#rows - 1)(#columns - 1)
goodness-of-fit: df = (#groups) − 1 − (# est. params)

- Model utility test:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = \frac{b_1}{\mathrm{SE}_{b_1}}, \quad \text{with} \quad df = n - 2$$

- F-test in ANOVA: $F = \dfrac{\mathrm{MSG}}{\mathrm{MSE}}$, where

$$\begin{aligned} df_{\text{numer}} &= (\#\text{ of groups}) - 1, \quad \text{and} \\ df_{\text{denom}} &= (\text{sample size}) - (\#\text{ of groups}) \end{aligned}$$

Probability

- Conditional probability: $\quad \mathrm{P}(B \mid A) = \dfrac{\mathrm{P}(A \text{ and } B)}{\mathrm{P}(A)}$
- Bayes' rule: $\quad \mathrm{P}(B \mid A) = \dfrac{\mathrm{P}(A \mid B)\,\mathrm{P}(B)}{\mathrm{P}(A)}$
- Total probability: $\quad \mathrm{P}(A) = \mathrm{P}(A \mid B)\,\mathrm{P}(B) + \mathrm{P}(A \mid B^c)\,\mathrm{P}(B^c)$

Miscellaneous

- Sample standard deviation $\quad s = \sqrt{\dfrac{1}{n-1}\sum_i (x_i - \bar{x})^2}$

Combinations of Random Variables

If X, Y are random variables, a,b are numbers, then

- $\mathrm{E}(aX) = a\,\mathrm{E}(X)$
- $\mathrm{E}(X \pm Y) = \mathrm{E}(X) \pm \mathrm{E}(Y)$
- $\mathrm{Var}(aX) = a^2 \cdot \mathrm{Var}(X)$, $\quad$ or $\quad \mathrm{SD}(aX) = |a| \cdot \mathrm{SD}(X)$
- Moreover, if X, Y are independent,

$\sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$, or $\sigma_{X\pm Y} = \sqrt{\sigma^2_X + \sigma^2_Y}$.

Least Squares Regression

The coefficients (from data) are given by

$$b_1 = r\frac{s_y}{s_x}, \qquad b_0 = \bar{y} - b\bar{x}$$