Distributions

- Sampling dist. for mean of sample of size n from population with mean μ, standard deviation σ

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

- Binomial: For $X \sim \mathsf{Binom}(n,p)$,

$$\mu_X = np\,, \qquad \sigma_X = \sqrt{np(1-p)}.$$

When $np \geq 10$ and $n(1-p) \geq 10$, then X has approx. dist. $\mathsf{Norm}(np, \sqrt{np(1-p)})$.

- Sampling dist. for sample proportion has

$$\mu_{\hat{p}} = p, \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

When $N > 20n$, $np \geq 10$, and $n(1-p) \geq 10$, then $\hat{p}$ has approx. dist. $\mathsf{Norm}(p, \sqrt{p(1-p)/n})$.

Probability

- conditional probability $\Pr(B \mid A) = \dfrac{\Pr(A \text{ and } B)}{\Pr(A)}$
- Bayes' rule $\Pr(B \mid A) = \dfrac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}$
- total probability
  $\Pr(A) = \Pr(A \mid B)\Pr(B) + \Pr(A \mid B^c)\Pr(B^c)$

Inference Procedures

- Level C Confidence Intervals (general):

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

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

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

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

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

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

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

As conservative estimate, take $df = \min(n_1, n_2) - 1$

- Chi-square test statistic:
  $\chi^2 = \sum \dfrac{[\text{(observed count) - (expected count)}]^2}{\text{expected count}}$
  contingency table: df = (#rows - 1)(#columns - 1)
  goodness-of-fit: df = (#groups) – 1 – (# est. params)
- Model utility test:

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

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

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

Miscellaneous

- Determine sample size, 1-proportion settings:
  To have margin of error no larger than a desired size M, take,

$$n \geq \left(\frac{z^*}{M}\right)^2 \tilde{p}(1-\tilde{p}),$$

where $\tilde{p}$ is an estimate of p (take $\tilde{p} = 0.5$ if no estimate is available)

Combinations of Random Variables

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

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

$$\sigma_{X\pm Y}^2 = \sigma_X^2 + \sigma_Y^2, \text{ or } \sigma_{X\pm Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}.$$

Least Squares Regression

The coefficients (from data) are given by

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