Random variables

- Generally, when $f(x)$ is the
  - pmf for (discrete) r.v. X: $\quad \mathrm{E}(X^j) = \sum_x x^j f(x)$
  - pdf for (continuous) r.v. X: $\quad \mathrm{E}(X^j) = \int_{-\infty}^{\infty} x^j f(x)\,dx$
- The variance is $\quad \mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$

Distributions

family	mean	variance
$X \sim \mathsf{Binom}(n,p)$	np	$np(1-p)$
$X \sim \mathsf{Geom}(p)$	$(1-p)/p$	$(1-p)/p^2$
$X \sim \mathsf{Pois}(\lambda)$	λ	λ
$X \sim \mathsf{Unif}(a,b)$	$\frac{1}{2}(a+b)$	$\frac{1}{12}(b-a)^2$
$X \sim \mathsf{Norm}(\mu,\sigma)$	μ	σ^2
$X \sim \mathsf{Exp}(\lambda)$	$1/\lambda$	$1/\lambda^2$

Sample statistics

For iid sample $X_1, \dots, X_n$ 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 $Y \sim \mathsf{Binom}(n,p)$. 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, $\quad \mathrm{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
  - z-test (when $\hat{p}$ approx. normal)

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

- 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 Welch $df = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$.

- Chi-square test statistic:

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

  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

  df_{numer} = (# of groups) - 1, and
  df_{denom} = (sample size) - (# of groups)

Miscellaneous

- Sample standard deviation $\quad s = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2}$
- 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: $\quad \mathrm{P}(A) = \mathrm{P}(A \mid B)\,\mathrm{P}(B) + \mathrm{P}(A \mid \overline{B})\,\mathrm{P}(\overline{B})$

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,

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

Least Squares Regression

The coefficients (from data) are given by

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