Statistics for Data Science - 2

Important results

Discrete random variables:

Distribution	PMF ($f_X(k)$)	CDF ($F_X(x)$)	$E[X]$	$\text{Var}(X)$
Uniform(A) $A = \{a, a+1, \ldots, b\}$	$\frac{1}{n}, \quad x = k$ $n = b - a + 1$ $k = a, a+1, \ldots, b$	$\begin{cases} 0 & x < 0 \\ \frac{k-a+1}{n} & k \leq x < k+1 \\ & k = a, a+1, \ldots, b-1, b \\ 1 & x \geq n \end{cases}$	$\frac{a+b}{2}$	$\frac{n^2-1}{12}$
Bernoulli(p)	$\begin{cases} p & x = 1 \\ 1-p & x = 0 \end{cases}$	$\begin{cases} 0 & x < 0 \\ 1-p & 0 \leq x < 1 \\ 1 & x \geq 1 \end{cases}$	p	$p(1-p)$
Binomial(n, p)	${}^nC_k p^k (1-p)^{n-k}$, $k = 0, 1, \ldots, n$	$\begin{cases} 0 & x < 0 \\ \sum_{i=0}^{k} {}^nC_i p^i (1-p)^{n-i} & k \leq x < k+1 \\ & k = 0, 1, \ldots, n \\ 1 & x \geq n \end{cases}$	np	$np(1-p)$
Geometric(p)	$(1-p)^{k-1}p$, $k = 1, \ldots, \infty$	$\begin{cases} 0 & x < 0 \\ 1-(1-p)^k & k \leq x < k+1 \\ & k = 1, \ldots, \infty \end{cases}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Poisson(λ)	$\frac{e^{-\lambda}\lambda^k}{k!}$, $k = 0, 1, \ldots, \infty$	$\begin{cases} 0 & x < 0 \\ e^{-\lambda} \sum_{i=0}^{k} \frac{\lambda^i}{i!} & k \leq x < k+1 \\ & k = 0, 1, \ldots, \infty \end{cases}$	λ	λ

Continuous random variables:

Distribution	PDF ($f_X(k)$)	CDF ($F_X(x)$)	$E[X]$	Var(X)
Uniform$[a,b]$	$\frac{1}{b-a}$, $a \leq x \leq b$	$\begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x \geq b \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Exp(λ)	$\lambda e^{-\lambda x}$, $x > 0$	$\begin{cases} 0 & x \leq 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Normal(μ, σ^2)	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$, $-\infty < x < \infty$	No closed form	μ	σ^2
Gamma(α, β)	$\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$, $x > 0$		$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$
Beta(α, β)	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ $0 < x < 1$		$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

1. **Markov's inequality:** Let X be a discrete random variable taking non-negative values with a finite mean μ. Then,
$$P(X \geq c) \leq \frac{\mu}{c}$$

2. **Chebyshev's inequality:** Let X be a discrete random variable with a finite mean μ and a finite variance σ^2. Then,
$$P(\mid X - \mu \mid \geq k\sigma) \leq \frac{1}{k^2}$$

3. **Weak Law of Large numbers:** Let $X_1, X_2, \ldots, X_n \sim$ iid X with $E[X] = \mu, \text{Var}(X) = \sigma^2$.
Define sample mean $\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$. Then,
$$P(|\overline{X} - \mu| > \delta) \leq \frac{\sigma^2}{n\delta^2}$$

4. **Using CLT to approximate probability:** Let $X_1, X_2, \ldots, X_n \sim$ iid X with $E[X] = \mu, \text{Var}(X) = \sigma^2$.
Define $Y = X_1 + X_2 + \ldots + X_n$. Then,
$$\frac{Y - n\mu}{\sqrt{n}\sigma} \approx \text{Normal}(0, 1).$$

- **Test for mean**
  Case (1): When population variance σ^2 is known (z-test)

Test	H_0	H_A	Test statistic	Rejection region
right-tailed	$\mu = \mu_0$	$\mu > \mu_0$	$T = \overline{X}$ $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\overline{X} > c$
left-tailed	$\mu = \mu_0$	$\mu < \mu_0$	$T = \overline{X}$ $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\overline{X} < c$
two-tailed	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \overline{X}$ $Z = \dfrac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$	$\lvert\overline{X} - \mu_0\rvert > c$

Case (2): When population variance σ^2 is unknown (t-test)

Test	H_0	H_A	Test statistic	Rejection region
right-tailed	$\mu = \mu_0$	$\mu > \mu_0$	$T = \overline{X}$ $t_{n-1} = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\overline{X} > c$
left-tailed	$\mu = \mu_0$	$\mu < \mu_0$	$T = \overline{X}$ $t_{n-1} = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\overline{X} < c$
two-tailed	$\mu = \mu_0$	$\mu \neq \mu_0$	$T = \overline{X}$ $t_{n-1} = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$	$\lvert\overline{X} - \mu_0\rvert > c$

- **χ^2-test for variance:**

Test	H_0	H_A	Test statistic	Rejection region
right-tailed	$\sigma = \sigma_0$	$\sigma > \sigma_0$	$T = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$S^2 > c^2$
left-tailed	$\sigma = \sigma_0$	$\sigma < \sigma_0$	$T = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$S^2 < c^2$
two-tailed	$\sigma = \sigma_0$	$\sigma \neq \sigma_0$	$T = \dfrac{(n-1)S^2}{\sigma_0^2} \sim \chi^2_{n-1}$	$S^2 > c^2$ where $\dfrac{\alpha}{2} = P(S^2 > c^2)$ or $S^2 < c^2$ where $\dfrac{\alpha}{2} = P(S^2 < c^2)$

- **Two samples z-test for means:**

Test	H_0	H_A	Test statistic	Rejection region
right-tailed	$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$T = \overline{X} - \overline{Y}$ $\overline{X} - \overline{Y} \sim \text{Normal}\left(0, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$ if H_0 is true	$\overline{X} - \overline{Y} > c$
left-tailed	$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	$T = \overline{Y} - \overline{X}$ $\overline{Y} - \overline{X} \sim \text{Normal}\left(0, \dfrac{\sigma_2^2}{n_2} + \dfrac{\sigma_1^2}{n_1}\right)$ if H_0 is true	$\overline{Y} - \overline{X} > c$
two-tailed	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$T = \overline{X} - \overline{Y}$ $\overline{X} - \overline{Y} \sim \text{Normal}\left(0, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$ if H_0 is true	$\lvert\overline{X} - \overline{Y}\rvert > c$

- **Two samples F-test for variances**

Test	H_0	H_A	Test statistic	Rejection region
one-tailed	$\sigma_1 = \sigma_2$	$\sigma_1 > \sigma_2$	$T = \dfrac{S_1^2}{S_2^2} \sim F_{(n_1-1, n_2-1)}$	$\dfrac{S_1^2}{S_2^2} > 1 + c$
one-tailed	$\sigma_1 = \sigma_2$	$\sigma_1 < \sigma_2$	$T = \dfrac{S_1^2}{S_2^2} \sim F_{(n_1-1, n_2-1)}$	$\dfrac{S_1^2}{S_2^2} < 1 - c$
two-tailed	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	$T = \dfrac{S_1^2}{S_2^2} \sim F_{(n_1-1, n_2-1)}$	$\dfrac{S_1^2}{S_2^2} > 1 + c_R$ where $\dfrac{\alpha}{2} = P(T > 1 + c_R)$ or $\dfrac{S_1^2}{S_2^2} < 1 - c_L$ where $\dfrac{\alpha}{2} = P(T < 1 - c_L)$

- **χ^2-test for goodness of fit:**
  H_0 : Samples are i.i.d X, $\quad H_A$: Samples are not i.i.d X

  Test statistic: $T = \sum_{i=1}^{k} \frac{(y_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} \sim \chi^2_{k-1}$

  Test: Reject H_0 if $T > c$.

- **Test for independence:**
  H_0 : Joint PMF is product of marginals, H_A : Joint PMF is not product of marginals

  Test statistic: $T = \sum_{i,j} \frac{(y_{ij} - np_{ij})^2}{np_{ij}} = \sum_{i=1}^{k} \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}} \sim \chi^2_{dof}$

  where $dof =$ (number of rows-1) $\times$ (number of columns-1)
  $y_{ij} =$ product of marginals for (i, j)
  $np_{ij} =$ expected, if independent

  Test: Reject H_0 if $T > c$.