SCIM122
AY24/25 SEM 2
skibidi

Probability

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$P(A \cap B) = P(A)P(B \mid A)$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A')}$$

Discrete distributions

For a discrete random variable X taking values x_i with probabilities $P(X = x_i)$

Expectation (mean): $E(X) = \mu = \sum x_i\, P(X = x_i)$

Variance: $Var(X) = \sigma^2 = \sum (x_i - \mu)^2\, P(X = x_i) = \sum x_i^2\, P(X = x_i) - \mu^2$

For a function $g(X)$: $E(g(X)) = \sum g(x_i)\, P(X = x_i)$

Random Variables

Types of Random Variables Binomial: Random variable with parameters n and p
$Y = X_1 + X_2 + \cdots + X_n$ are independent.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

P(k successes from n independent trials each with probability of p success).
Example: number of red balls out of n balls drawn with replacement.

$$\mathbb{E}(Y) = np, \quad \mathrm{Var}(Y) = np(1-p)$$

Negative Binomial: $X =$ number of trials until k successes are obtained.

$$P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}, \quad \text{for } x = k, k+1, \ldots$$

Geometric: $X =$ number of trials until a success is obtained.where k is the number of trials needed

$$P(X = k) = (1-p)^{k-1} p$$

Hypergeometric: $X =$ number of trials until success, without replacement.

$$P(X = x) = \frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}}, \quad \text{for possible } x$$

Poisson distribution:

$$P(X = x) = \frac{e^{-\mu}\mu^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots$$

Variance

Definition. Let X be a random variable with probability distribution $f(x)$ and mean μ.
The **variance** of X is

$$\sigma_X^2 = E[(X - \mu_X)^2] = \sum_x (x - \mu_X)^2 f(x) (X \text{ is discrete}),$$

$$\sigma_X^2 = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x)\, dx (X \text{ is continuous}).$$

The **standard deviation** of X is the (non-negative) square root of the variance.

$$\sigma_X^2 = E(X^2) - \mu_X^2$$

If X is a random variable, and are constants, then

$$\sigma_{aX+b}^2 = \sigma_{aX}^2 = a^2\sigma_X^2$$

Continuous Uniform Distribution

Suppose that X is a continuous uniform random variable on the interval $[A, B]$. The density function of X is

$$f(x) = \begin{cases} \frac{1}{B-A}, & A \le X \le B, \\ 0, & \text{otherwise.} \end{cases}$$

Exponential Distribution

Suppose that X follows an exponential distribution with the density function

$$f(x) = \begin{cases} \frac{1}{\beta} e^{-x/\beta}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

where β is a positive constant.

Joint Probability Distributions

Definition. The function $f(x, y)$ is a **joint probability distribution** (or **probability mass function**) of discrete random variables X and Y if:

1. $f(x, y) \ge 0$ for all (x, y),
2. $\sum_x \sum_y f(x, y) = 1$,
3. $P(X = x, Y = y) = f(x, y)$.

For any region A in the xy-plane,

$$P((X, Y) \in A) = \sum_{(x,y) \in A} f(x, y).$$

Definition. The function $f(x, y)$ is a **joint density function** of continuous random variables X and Y if:

1. $f(x, y) \ge 0$ for all (x, y),
2. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$,
3. $P((X, Y) \in A) = \iint_A f(x, y)\, dx\, dy$, for any region A in the xy-plane.

Estimating Proportion

Estimating a proportion

An (approximate) $100(1 - \alpha)\%$ confidence interval for the proportion p is

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $\hat{p}$ = the proportion of successes in a random sample of size n, and $\hat{q} = 1 - \hat{p}$

- sample of size **n**
- number of successes **X**
- proportion of success $\hat{p} = \frac{X}{n}$
- Standard Error,SE $= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$
- Confidence interval = α
- find $z_{\alpha/2}$ value of $\alpha/2$ in table

Estimating the difference between two proportions

An (approximate) $100(1 - \alpha)\%$ confidence interval for the difference of two proportions $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

Estimating variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$$

Unbiased estimator

The sample mean is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The sample variance is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Chi-square distribution

A chi-squared distribution is a continuous probability distribution whose density function is

$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

where the parameter ν is called the degrees of freedom (ν is a positive integer),
and Γ is the Gamma function, i.e.

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x}\, dx \quad \text{for } \alpha > 0.$$

Constructing a confidence interval for a variance

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

has the chi-squared distribution with $\nu = n - 1$ degrees of freedom.

$$P\left(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}\right) = 1 - \alpha$$

If s_1^2 and s_2^2 are the variances of independent random samples of size n_1 and n_2, respectively, from normal populations, then a $100(1 - \alpha)\%$ confidence interval for $\frac{\sigma_1^2}{\sigma_2^2}$ is

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot f_{\alpha/2}(v_2, v_1)$$

where $f_{\alpha/2}(v_1, v_2)$ is the critical value of the F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom, with area on the right tail $\alpha/2$
(and $f_{\alpha/2}(v_2, v_1)$ is defined similarly).

Maximum likelihood estimation

Item 1
non-defective

Item 2
non-defective

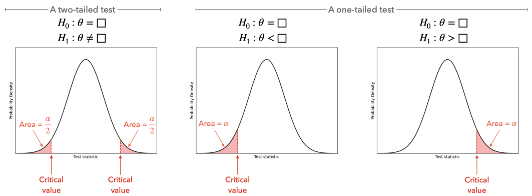
Item 3
defective

$$p \times p \times (1 - p) = p^2 - p^3.$$

The **level of significance** (α) = the probability of rejecting H_0 when it is true.

= the probability of making a **Type I error**.

The **critical region** (where we decide to reject H_0) depends on α and H_a.



Hypothesis testing for two variances

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2}$$

has the F-distribution with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom. The critical region is $f_\alpha(\nu_1, \nu_2)$ in F-distribution table, α is critical value

Test for independence

expected frequency,$e_i = \frac{\text{(column total)} \times \text{(row total)}}{\text{grand total}}$

Test statistic: $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i}$

Observed frequency,o_i,from table in question

Confidence Interval for Difference in Proportions

A marketing research is investigating customers' preferences between two brands of a particular product. In two independent samples, 180 out of 400 people have heard about brand A, while 210 out of 500 have heard about brand B. Construct a 95% confidence interval for the difference:$p_A - p_B$, where p_A and p_B are the proportions of customers who have heard about brand A and brand B, respectively.

$$\hat{p}_A = \frac{180}{400} = 0.45, \hat{p}_B = \frac{210}{500} = 0.42$$

$$z^* = 1.96$$

$$\text{CI} = (\hat{p}_A - \hat{p}_B) \pm z^* \cdot \sqrt{\frac{\hat{p}_A(1 - \hat{p}_A)}{n_A} + \frac{\hat{p}_B(1 - \hat{p}_B)}{n_B}}$$

$$= 0.03 \pm 1.96 \cdot \sqrt{\frac{0.45 \cdot 0.55}{400} + \frac{0.42 \cdot 0.58}{500}}$$

$$= 0.03 \pm 1.96(0.03326) \Rightarrow \text{CI} = (-0.0352,\ 0.0952)$$

A manufacturing company wants to ensure consistency in the thickness of its glass panels. Let $X_1, X_2, \ldots, X_{30}$ be the thickness of 30 panels in a random sample. Suppose that:

$$\sum_{i=1}^{30} X_i = 110.6 \text{ mm}, \quad \sum_{i=1}^{30} X_i^2 = 478.24 \text{ mm}^2$$

Calculate the variance of this sample.

$$s^2 = \frac{1}{n-1}\left(\sum X_i^2 - \frac{(\sum X_i)^2}{n}\right)$$

$$= \frac{1}{29}\left(478.24 - \frac{(110.6)^2}{30}\right)$$

$$= \frac{1}{29}(478.24 - 407.75) = 2.4309$$

In country A, each person has a 25% chance of having blood type O. In country B, each person has a 30% chance of having blood type O. 30 people from country A are randomly chosen, and let X be the number of people in this group who have blood type O.
25 people from country B are randomly chosen, and let Y be the number of people in this group who have blood type O. **Which random variable has a greater variance, X or Y?** Since both X and Y follow Binomial distributions:

$$\mathrm{Var}(X) = 30 \cdot 0.25 \cdot 0.75 = 5.625$$

$$\mathrm{Var}(Y) = 25 \cdot 0.30 \cdot 0.70 = 5.25$$

$\Rightarrow$ So, X has a greater variance.