STATISTICS

Introduction

Statistics: analyze, interpret, transform data into information

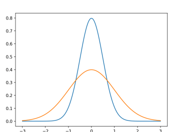
- **Descriptive statistics**: organize, summarize and present
- **Inferential statistics**: use sample data to estimate
  - Sample stats ($\bar{x}$, s), Population parameters (μ, σ)
- **Graphical methods**: histograms, boxplots
- **Summary statistics**: Percentile, Median, Quartile, IQR
- If randomly assigned, not much systematic difference
- Measure center: **Mean** ($\bar{x}$), **Median**, (expectation)
- Measure variability: **Range**, **IQR**, s^2, (risk/uncertainty)
- Mean is more affected by outliers than Median
  - Median for highly skewed; Mean for symmetrical

> Skewness:
> - Symmetric if Median = Mean
> - Positively (right) skew if Median < Mean
> - Negatively (left) skew if Median > Mean
>
> Median % away from Mean:
>
> $$\frac{\text{Median}}{\text{Mean}}$$
>
> - No material skew: Median in 5% of Mean
> - Mean-modest skew: Median in 5% $\sim$ 20% of Mean
> - Mean-high skew: Median in 20% of Mean

- Sample variance s^2:

$$s^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

$n-1$ correction: two estimation ($\bar{x} \to \mu$, $s^2 \to \sigma^2$)



s^2: unbiased estimator for σ^2
With smaller s, estimation for μ more accurate

- **Empirical rule**: 68%, 95%, 99.7% of data within 1, 2, 3 s.d. for normal distribution; Over 3 s.d. is suspicious
- Guesstimate $s = range/\sigma$

Probability

- **Rule**

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

- Mutually exclusive, $P(A \cup B) = P(A) + P(B)$
- Two events A and B are **independent** if:
  - $P(A|B) = P(A|\bar{B}) = P(A)$
  - $P(B|A) = P(B|\bar{A}) = P(B)$
  - $P(A \cap B) = P(A) \cdot P(B)$
- Mutually exclusive events are NOT independent
- **Law of Total Probability**

$$P(A) = \Sigma P(A|B_i) \cdot P(B_i)$$

- **Bayes's Rule**

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i) \cdot P(A_i)}{P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n)}$$

Random Variables (RVs)

- **Binomial RV**

$$\boxed{\mu = np,\ \sigma^2 = npq}$$

  - When $p = 0.5$, largest variability
- Sample space: the set of all possible outcomes
- RV: a variable that assigns value to each outcome
- **Discrete RV**
  - PMF: $p(x) \geq 0$, $\forall x$, $\Sigma p(x) = 1$

$$\boxed{\mu = \Sigma x \cdot p(x),\ \sigma^2 = \Sigma(x-\mu)^2 \cdot p(x)}$$

- **Continuous RV**
  - PMF: $p(x) \geq 0$, $\forall x$, $\int p(x)\,dx = 1$, $p(c) = 0$
  - Probability is the area under curve

$$\boxed{\mu = \Sigma x \cdot p(x),\ \sigma^2 = \Sigma(x-\mu)^2 \cdot p(x)}$$

- **Normal Distribution**
  - Mean, Median, and Mode are equal

$$\boxed{\begin{array}{l} z = (x-\mu)/\sigma,\ \mu = 0,\ \sigma = 1 \\ x = \mu + z \cdot \sigma \end{array}}$$

- **Uniform Distribution**

$$\boxed{\begin{array}{l} f(x) = 1/(d-c),\ \mu = (c+d)/2,\ \sigma = (d-c)/\sqrt{12} \\ P(a < x < b) = (b-a)/(d-c) \end{array}}$$

- **Population Average Treatment Effect (PATE)** measures average difference in outcomes between treated group and control group across the population

$$PATE = E(Y|T) - E(Y|T^c)$$

- **Randomization creates balance** between treatment groups, groups are expected to be similar on average w.r.t. both observed and unobserved characteristics

Central Limit Theorem

- **Central Limit Theorem (CLT)**: the sampling distribution of the sample mean $\bar{x}$ of a sufficiently large sample ($N \geq 30$) will approximate a normal distribution.
- Make probabilistic statements about $\bar{x}$'s relation to μ

$$\mu_{\bar{x}} = \mu, \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

More data, larger n, more reliable estimates of μ

Confidence Interval for population mean μ

- **z-score** $\rightarrow$ how many s.d. a data is away from the mean
- c $\rightarrow$ confidence level, the probability that a range contains the true population parameter (usually 95%)
- α $\rightarrow$ significance level, the likelihood that the true population parameter is out of the range, $\alpha = 1 - c$
- E $\rightarrow$ error bound/margin or error

Confidence interval of z-score

$$\bar{x} \pm E = \bar{x} \pm \left(z_{c/2} \cdot \frac{s}{\sqrt{N}}\right)$$

- Larger N (more data) $\rightarrow$ interval$\downarrow$
- Larger c (relaxed) $\rightarrow$ interval$\uparrow$

Given **sample** $\bar{x}$ and s, estimate **population** μ: there is 95% chance that μ is between ... and ...

$$N = (\frac{z_{c/2} \cdot s}{E})^2$$

- **t-score**: When $s \rightarrow \sigma$, s is a sample statistic

Confidence interval of t-score

$$\bar{x} \pm t^{df}_{c/2}(\frac{s}{\sqrt{N}})$$

where $df = N - 1$

- For large $N > 30$, t-score converges to z-score
- For small $N < 30$, t-score is recommended, which accounts for the additional variability (though CLT may not hold)

Confidence Interval for population proportion p

- $E(X) = \mu = 1p + 0(1-p) = p$
- $Var(X) = \sigma^2 = (1-p)^2 p + (0-p)^2 (1-p) = p(1-p)$
- **CLT** applies since p is sample average, $E(X) = p$
  - Rule of thumb: $np > 15$, $n(1-p) > 15$
  - CLT may not hold when p is very close to 0 or 1

Confidence interval for p

$$\hat{p} \pm Z_{c/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

A guess of $p = 0.5$ will yield the largest N

$$N = (\frac{Z_{c/2}}{E})^2 \cdot p(1-p)$$

Hypothesis Testings

- Assumptions: 1) **CLT** holds; 2) Randomized

H_0: null hypothesis assumed to be true ($\neq, >, <$)
H_1: alternative hypothesis ($=, \leq, \geq$)

$$z/t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{N}}}$$

Not used for population proportions (no $s \rightarrow \sigma$)

	Do not reject H0	Reject H0
H0 True	correct	**Type I** (α)
H0 False	**Type II** (β)	correct

- For fixed sample size n, decreasing α increases β
- Increasing sample size n decreases β

Difference between two means

CI for two means (t-score):

$$(\bar{x}_1 - \bar{x}_2) \pm t^{df}_{\alpha/2} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $df = n_1 + n_2 - 2$

CI for two proportions (z-score):

$$(\hat{p_1} - \hat{p_2}) \pm Z_{c/2} \cdot \sqrt{\hat{p}(1-\hat{p}) \cdot (\frac{1}{n_1} + \frac{1}{n_2})}$$

where $\hat{p} = (\hat{p_1} n_1 + \hat{p_2} n_2)/(n_1 + n_2)$

Regression

- $y_i = \beta_0 + \beta_1 \cdot X_i + \epsilon_i$ (β_0, β_1 deterministic, ϵ_i random)
  - Assume relationship can be approximated linearly
  - Assume random error for each data point is drawn independently from a normal distribution with mean 0 and s.d. σ_ϵ : $e_i \sim N(0, \sigma_\epsilon)$, $P(e|x) = P(e)$
- Minimize SSE: $\sum_i (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$
- $\hat{\beta}_1 = \frac{ss_{xy}}{ss_{xx}}$, $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- Residual standard error, $s_\epsilon = \sqrt{\Sigma \frac{(y_i - \hat{y})^2}{n-2}}$
- $s_{\hat{\beta}_1}$ measures the precision of $\hat{\beta}_1$ as an estimate for β_1

$$s_{\hat{\beta}} = s_\epsilon / \sqrt{ss_{xx}}, ss_{xx} = \sum^n (x_i - \bar{x})^2$$

- Linear Regression:
  - $y = \beta_0 + \beta_1 \cdot x$, (y dependent and x independent)
  - $y = \beta_0 x^{\beta_1}$, $logy = \beta_0 + \beta_1 \cdot logx$
- Testing significance of coefficients:

$$t = \frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}}, df = n - 2$$

  Reject: $t > t_{\alpha/2}, n-2$ (2-tailed), $t > t_{\alpha, n-2}$ (1-tailed)
- Confidence interval on coefficients:

$$\hat{\beta} \pm E = \hat{\beta} \pm t^{df}_{\alpha/2} \cdot s_{\hat{\beta}}$$

- **Coefficient of Determination** (R^2): proportion/percentage of variation in y that is "explained" by regression
  - $R^2 = 1$: perfect linear relationship
  - $0 < R^2 < 1$: not perfect linear relationship
  - $R^2 = 0$: no linear relationship
- Multiple Regression: $df = n - k + 1$, where k is the number of independent variables
- **Correlation** r: the association between x_1 and x_2 $P(\epsilon_i|x) = P(\epsilon)$ implies $r = 0$
- Randomization reduces the risk of omitted variable bias
- Avoid Omitted Variable Bias: 1) Add more regressors; 2) Randomization; 3) Instrumental Variable