Statistical Methods I (Lecture 5)

Outline:

- Parameter, Statistics, Estimate
- Point estimate: methods of moments
- Interval estimate: sampling distribution
- Summary of hypothesis testing
  - One-sample test
  - Two-sample test

Population vs. Sample

- Population attributes
  - $X, Y, \ldots$ (capital letters): **random variable** following some probability model or data generating process
  - $\theta, \mu, \sigma, \ldots$(Greek letters): intrinsic **population parameters** in some probability model
- Sample attributes
  - $x_1, x_2, \bar{x}, s, \ldots$ (small letters): (a function of) the **observed** values/outcome of r.v.'s in a particular data set.
  - $\hat{\theta}, \hat{\mu}, \hat{\sigma}, \ldots$("hat"): **estimated parameter/estimate** from a particular data set.
- Example: A survey conducted by a research in art education found that, 17% of those surveyed, had taken one course in dance in their life.

  Q: Is the number 17% a sample attribute or a population attribute?

Sample and Statistics

- Let $(X_1, \ldots, X_n)$ be a random sample of size n. Any random variable $T = f(X_1, \ldots, X_n)$ as a function of $(X_1, \ldots, X_n)$ is called a **statistic**.
  - If we treat each X_i as a random variable, T is called an **estimator**.
  - If we plug X_i by the observed value from a particular sample, T is called an **estimate**.
- Dance Survey Problem: Is the 17% an estimate, estimator, or parameter? What is the statistics in this setting?
- Example
  - The sample mean, defined by $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, is a **statistics**.
  - The sample variance, defined by $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$, is a **statistics.**
  - Why capital letter?
- Note: A statistic/estimator/estimate cannot involve any unknown parameter in its expression. For example, $\bar{X} - \mu$ is not a statistic if the population mean μ is unknown.

Sample and Statistics

- Key: A statistic (a function from sample) can be viewed as a **random variable** varying from sample to sample.
- How to infer the **population attributes (parameter)** from the sample statistic?
- Point estimate
  - Objective: obtain an "good" guess of a population parameter from a sample statistic.
  - Methods: methods of moment, least sum of squares, MLE, etc.
- Interval estimate.
  - Objective: obtain an "good" interval in which the population parameter will most likely lie on.
  - Methods: Distribution of sample statistics.

Point estimation (Method of moments)

- Use the data you have to calculate **sample moments** or **centered sample moments**.
- To fit a certain distribution, use relation to moments formula:
  - Option 1:
    $$\mathbb{E}(X^k) = \hat{\mu}_k \equiv \frac{1}{n}\sum_{i=1}^{n} x_i^k$$
    where $\mathbb{E}(X^k)$ is k-th population moments and $\hat{\mu}_k$ is k-th sample moment (**from data**);
  - Option 2:
    $$\mathbb{E}\left[(X - \mathbb{E}X)^k\right] = \hat{\mu}'_k \equiv \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$
    where $\mathbb{E}\left[(X - \mathbb{E}X)^k\right]$ is k-th centered population moments and $\hat{\mu}'_k$ is k-th centered sample moment.

Example: Method of Moments

- Suppose Michale recorded the temperatures (^oF) at noon for recent 10 days

50	60	45	52	67	76	80	68	75	82

- Sample Mean: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 65.5$$

  Sample Variance: $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 173.83$$

  So 2^{nd} centered sample moment:

$$\hat{\mu}'_2 = \frac{n-1}{n}s^2 = 173.83 \times 9/10 = 156.45.$$

- Note: 2^{nd} centered sample moment μ'_2 is different from sample variance s^2.

Example: Method of Moments

Temperature:	50	60	45	52	67	76	80	68	75	82

- Model 1: Suppose we want to fit an i.i.d. uniform $U(a, b)$ model

$$f_X(x) = \frac{1}{b-a} \quad a \leq x \leq b.$$

i.e. what is the estimate of a and b?

Remember $E(X) = \frac{(a+b)}{2}$, and $Var(X) = \frac{(b-a)^2}{12}$. Now use "relation to moment" formula

$$\frac{(a+b)}{2} = E(X) = \bar{x} = 65.6,$$

$$\frac{(b-a)^2}{12} = Var(X) = \hat{\mu}_2 = 156.45.$$

Therefore we have $\hat{a} = 43.93, \hat{b} = 87.26$.

Example

Temperature:	50	60	45	52	67	76	80	68	75	82

- Model 2: Suppose we want to fit with an i.i.d. $N(\mu, \sigma)$ model

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \left[-\frac{1}{2} \left(\frac{x-\mu}{\sigma}^2 \right) \right].$$

  i.e. what is the estimate of μ and σ?

Remember $E(X) = \mu$, and $Var(X) = \sigma^2$. Now use "relation to moment" formula

$$\mu = E(X) = \bar{x} = 65.6,$$

$$\sigma^2 = Var(X) = \hat{\mu}_2 = 156.5.$$

Solving the above gives $\hat{\mu} = 65.6$ and $\hat{\sigma} = 12.6$.

Generalization: method of moments

In general, estimate m parameters, need m sample moments

Exponential-(λ) Distribution

- Sample Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Population Mean

$$E(X) = 1/\lambda$$

- Parameter Estimate:

$$\hat{\lambda} = 1/\bar{x}$$

Possion(λ) Distribution

- Sample Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Population Mean

$$E(X) = \lambda$$

- Parameter Estimate:

$$\hat{\lambda} = \bar{x}$$

Generalization: method of moments

Aren't there other estimators?

Exponential-(λ) Distribution

- 2^{nd} centered sample moment

$$\mu_2' \equiv \frac{n-1}{n} s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - x)^2$$

- Population Variance

$$Var(X) = 1/\lambda^2$$

- Parameter Estimate:

$$\hat{\lambda} = \sqrt{1/\mu_2'} = \sqrt{ns^2/(n-1)}$$

Poisson(λ) Distribution

- 2^{nd} centered sample moment

$$\mu_2' \equiv \frac{n-1}{n} s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - x)^2$$

- Population Variance

$$Var(X) = \lambda$$

- Parameter Estimate:

$$\hat{\lambda} = \mu_2' = \frac{n-1}{n} s^2$$

Method of Moments

- Advantages
  - Simple to generate
  - Asymptotically normal (tends to normal when sample size n is large)
- Disadvantages:
  - Inconsistent results (more than one estimator equation)
  - Do not know how close the estimate is from parameter of interest.

Sampling Distribution

Sampling Distribution: the probability distribution of a given random-sample-based statistic.

Let $(X_1, X_2, \ldots, X_n)$ be an **i.i.d.** sample drawn from $N(\mu, \sigma^2)$.

Parameter (Population)	Estimator (Sample)	Distribution (do we need $n \to \infty$?)	Property
mean μ	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1)$	Unbiased
variance σ^2	$\hat{\sigma}^2 (= S^2) = \frac{1}{(n-1)}\sum_{i=1}^{n}(X_i - \bar{X})^2$	$\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \to \chi^2(n-1)$	Unbiased

- An estimator $\hat{\theta}$ for a parameter θ is called **unbiased estimator** if

$$\mathbb{E}(\hat{\theta}) = \theta$$

- Overestimate: $\mathbb{E}(\hat{\theta}) > \theta$; Underestiamte: $\mathbb{E}(\hat{\theta}) < \theta$.

Sampling Distribution of Estimators/Statistics

In general, let $\hat{\theta}$ be an estimator. How to find its bias?

- Express the estimator $\hat{\theta}$ as a function of sample $(X_1, ..., X_n)$. (hint: don't plug in the numerical value associated with a particular sample.)
- Treat each component $X_1, ..., X_n$ as **a random variable** with the **population** distribution.
- Use the properties of expectation (e.g., linearity) to calculate the expectation of $\hat{\theta}$.
- Compare $\mathbb{E}(\hat{\theta})$ with the real population parameter θ.
- In-class example

Sampling distribution

Example: Temperature Problem

- What Michael observed is a sample of the temperatures for 10 days.

50	60	45	52	67	76	80	68	75	82

  His estimate of population mean is

$$\hat{\mu} = \bar{x} = 65.6.$$

- Suppose Army also recorded the temperatures at the same location for recent 10 days

53	61	46	52	66	78	78	69	75	81

  What is her estimate for population mean?

$$\hat{\mu} = \bar{x} = 65.9.$$

- Why different $\hat{\mu}$? Who is right?

Distribution of Test Statistics

Let $(X_1, X_2, ..., X_n)$ be an i.i.d. sample drawn from a population $N(\mu, \sigma^2)$.

- If μ is unknown, σ is known, then
  - Sample Mean:
  $$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$
  - Sample Variance:
  $$\frac{(n-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-1}$$
- If both μ and σ are unknown, then

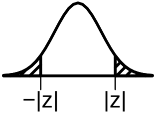
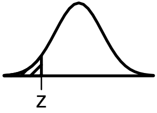
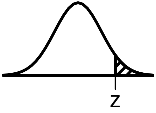
$$T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim T_{n-1}$$

.

Summary: Hypothesis Testing on Population Mean

If σ is known, z-statistics: $z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$

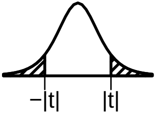
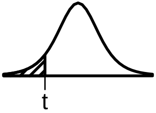
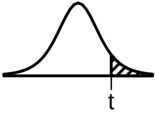
significance level	α	0.1	0.05	0.01
	$1-\alpha$	90%	95%	99%
N(0,1)	$z^*_{\alpha/2}$	1.64	1.96	2.58

	Two-Sided	Lower One-Sided	Upper One-Sided
H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
Test Statistic	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$		
P-value	$P(\lvert\text{Norm}(0,1)\rvert > \lvert z\rvert)$ 	$P(\text{Norm}(0,1) < z)$ 	$P(\text{Norm}(0,1) > z)$ 
Accept H_0 w/ significance level α	$\lvert z\rvert < z^*_{\alpha/2}$ or equivalently $\lvert\bar{x} - \mu_0\rvert < z^*_{\alpha/2}\frac{\sigma}{\sqrt{n}}$	$z > -z^*_{\alpha}$ or equivalently $\bar{x} - \mu_0 > -z^*_{\alpha}\frac{\sigma}{\sqrt{n}}$	$z < z^*_{\alpha}$ or equivalently $\bar{x} - \mu_0 < z^*_{\alpha}\frac{\sigma}{\sqrt{n}}$

Summary: Hypothesis Testing on Population Mean

If σ is unknown, t-statistics: $t = \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}} \sim T_{n-1}$

significance level	α	0.1	0.05	0.01
	$1-\alpha$	90%	95%	99%
T_{n-1}	$t^*_{n-1,\alpha/2}$		qt(α/2, n-1)	

	Two-Sided	Lower One-Sided	Upper One-Sided
H_0	$\mu = \mu_0$	$\mu = \mu_0$	$\mu = \mu_0$
H_1	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
Test Statistic		$t = \frac{\bar{x}-\mu_0}{\hat{\sigma}/\sqrt{n}}$	
P-value	$P(\lvert T_{n-1}\rvert > \lvert t\rvert)$ 	$P(T_{n-1} < t)$ 	$P(T_{n-1} > t)$ 
Accept H_0 w/ significance level α	$\lvert t\rvert < t^*_{n-1,\alpha/2}$ or equivalently $\lvert\bar{x}-\mu_0\rvert < t^*_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}$	$t > -t^*_{n-1,\alpha}$ or equivalently $\bar{x}-\mu_0 > -t^*_{n-1,\alpha}\frac{\hat{\sigma}}{\sqrt{n}}$	$t < t^*_{n-1,\alpha}$ or equivalently $\bar{x}-\mu_0 < t^*_{n-1,\alpha}\frac{\hat{\sigma}}{\sqrt{n}}$

Confidence Interval (Variance Known)

Parameter of interest: population mean μ

- When σ^2 known: z-test
  - Statistics
$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$
  - 95% **Confidence Interval (CI)**
$$\mu \in (\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \quad , \quad \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$
  - $(1-\alpha)$-**Confidence Interval (CI)**:
$$\mu \in (\bar{x} - z^*_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad , \quad \bar{x} + z^*_{\alpha/2}\frac{\sigma}{\sqrt{n}})$$
  - $z^*_{\alpha/2}$ is called critical value at level $\alpha/2$.

significance level	α	0.1	0.05	0.01
	$1-\alpha$	90%	95%	90%
N(0,1)	$z^*_{\alpha/2}$	1.64	1.96	2.58

Confidence Interval (Variance Unknown)

Parameter of interest: population mean μ

- When σ^2 is unknown: t-test
  - Statistics
    $$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}}$$
  - $(1-\alpha)$-**Confidence Interval (CI)**
    $$\mu \in (\bar{x} - t^*_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}} \quad , \quad \bar{x} + t^*_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}})$$
  - $t^*_{n-1,\alpha/2}$ is called critical value at level $\alpha/2$.

    In R: qt(...,df=n-1).

Margin of Error & Sample Size & Confidence Level

$$\underbrace{\bar{x}}_{\text{estimate}} \pm \underbrace{z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}_{\text{margin of error}} \quad \text{or} \quad \underbrace{\bar{x}}_{\text{estimate}} \pm \underbrace{t_{n-1,\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}}_{\text{margin of error}}$$

The size of the margin of error can be reduced if

- confidence level is smaller (e.g. 95% $\to$ 90%);
- sample size n is larger;
- or if σ is smaller

We usually prefer **shorter** Confidence Interval.

Duality of Confidence Intervals and Hypothesis Tests

$$\boxed{\text{In a two sided test, } H_0 : \mu = \mu_0 \text{ is not rejected at level } \alpha}$$

if and only if

$$\boxed{\mu_0 \text{ is in the } (1-\alpha) \text{ CI for } \mu}$$

Proof: In a two sided z-test, $H_0 : \mu = \mu_0$ is not rejected if

$$\begin{aligned}
|\bar{x} - \mu_0| \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}} &\iff -z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \bar{x} - \mu_0 \le z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \\
&\iff \bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu_0 \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}
\end{aligned}$$

Confidence Interval

In general, $(1-\alpha)$-CI for population parameter θ can be calculated from the test statistic for θ.

- Find the test statistic for θ and its null distribution;
- Find the critical value at level $\alpha/2$ (if two sided test) based on null distribution, say $c^*_{\alpha/2}$;
- Write the $(1-\alpha)$-CI in the form of

$$\text{estimate} \pm \text{margin of error}$$

  where the margin of error usually is the $c^*_{\alpha/2}\times$denominator in test statistics.

Assumption for T-test vs. Z-test

- In either case, observations must be i.i.d.
- Z-test: σ known.
- T-test: σ unknown, thus replace σ by sample variance $\hat{\sigma}$.
- Give similar results when sample size is large.
- The population distributions should be normal if n is low, if however $n > 30$ normality assumption is not required.

Comparison of Two Population Means: Paired T Test

- Parameter of interest: $\mu_1 - \mu_2$
- Data: $D_1 = y_1 - y_2, \ldots, D_n = y_1 - y_n$
- Paired two-sample inference:
  - Hypothesis testing $H_0 : \mu_D = \mu_D^0$

$$T = \frac{\bar{D} - \mu_D^0}{S_D/\sqrt{n}} \sim T_{n-1}, \text{ where } S_D = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2},$$

  - $(1-\alpha)$ CI for $\mu_D = \mu_1 - \mu_2$:

$$\bar{d} \pm t_{n-1,\alpha/2}\frac{s_D}{\sqrt{n}}$$

Comparison of Two Population Means: Independent Two Sample T Test

- Independent two-sample inference assuming $\sigma_1^2 = \sigma_2^2$:
  - Hypothesis testing $H_0 : \mu_1 - \mu_2 = \mu_D^0$

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - \mu_D^0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim T_{n_1+n_2-2},$$

    where $S_p^2 = n_1 + n_2 - 2\sum_{i=1}^{n_1}(Y_{1i} - \bar{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \bar{Y}_2)$.
  - $(1-\alpha)$ CI for $\mu_1 - \mu_2$:

$$\bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2,\alpha/2}\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Quiz

- A camera is recommended to be sold at price of μ, with a standard deviation of \$12. In a sample of 50 randomly selected stores, the average price is \$194.
  - Find the 95%-CI for average price of camera.
  - Based on the observation, do we have strong evidence to claim that the recommended price is set as $\mu = \$190$ with 0.05 type 1 error?
- A manufacturer claims that his tires last 40,000 miles on average. A test on 25 tires reveals that the mean life of a tire is 39,750 miles, with a sample standard deviation of 387 miles.
  - Find the 95%-CI for the average lifetime of the tire.
  - Based on the observation, can we reject the the manufacturer's claim with 0.05 type 1 error?

Answer to Problem 1:

- Let X_i denote the price of the camera in store i, where $i = 1, \ldots, n$.
- Assumption: $\{X_i\}$ is an i.i.d. sample with $\mu = \mathbb{E}(X_i)$ and $\sigma^2 = \mathrm{Var}(X_i)$. Since the sample size $n \geq 30$, the normality assumption can be relaxed.
- Since the true variance is known, we consider the Z-statistics:
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$
- The 95%-CI for μ is
$$\mu \in [\bar{x} + \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}].$$
Note that $\bar{x} = 194$, $\sigma = 12$, $z_{\alpha/2} = 1.96$. Therefore the 95%-CI is $[190.7, 197.3]$.
- The claim $\mu = \$190$ is outside the 95%-CI $[190.7, 197.3]$. Due to the duality of CI and hypothesis testing, we reject the claim $H_0 : \mu = 190$ with 0.05 type 1 error.

Answer to Problem 2:

- Let X_i denote the lifetime of the tire i, where $i = 1, \ldots, n$.
- Assumption: $\{X_i\}$ is an i.i.d. sample, where $X_i \sim N(\mu, \sigma^2)$.
- Since the true variance is unknown, we consider the T-statistics:

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim T_{n-1}, \quad \text{where } \hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- The 95%-CI for μ is

$$\mu \in [\bar{x} + \frac{\hat{\sigma}}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{x} - \frac{\hat{\sigma}}{\sqrt{n}} t_{n-1,\alpha/2}].$$

  Note that $\bar{x} = 39750$ (miles), $\hat{\sigma} = 387$ (miles), $t_{24,0.025} = 2.064$. Therefore, the 95%-CI for μ is $[39590,\ 39913]$.

- Since the $\mu = 40000$ (miles) is outside the 95%-CI, we reject the claim $H_0 : \mu = 40000$ (miles) with 0.05 type 1 error.