① Estimation

(1.1) Statistic: statistic is a RV function of observations $X_1, \dots, X_n$ usually used to estimate some unknown parameter from the underlying probability distribution of the X_i's.

Ex: $\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$ (sample mean) $\quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$ (sample variance)

Let $X_1, \dots, X_n$ be iid RV's and let $T(\mathbf{X}) \equiv T(X_1, \dots, X_n)$ be a statistic based on the X_i's (it's also RV, since if we had different samples, we'd expect to get different values for the statistic).

If we use $T(\mathbf{X})$ to estimate some unknown parameter θ, $T(\mathbf{X})$ is called a point estimator for θ.

Ex: $\bar{X}$ is usually a point estimator for the mean $\mu = E[X]$

S^2 is often a point estimator for the variance $\sigma^2 = var(X_i)$

Desired Properties for any point estimator

- Unbiasedness ($E[T(X)] = \theta$)
- low variance ("efficiency")

Think: $E[(T(X)-\theta)]^2 = var(T(X)) + [Bias(T(X))]^2 \quad \left\{\begin{array}{l}\lim_{n\to\infty} Var(T(X)) = 0 \\ \lim_{n\to\infty} Bias(T(X)) = 0\end{array}\right\} T(X)$ is consistent

- $\bar{X}$ is always unbiased for μ : $E[\bar{X}] = E\left[\sum_{i=1}^{n} X_i/n\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n} n\mu = \mu$
- S^2 is always unbiased for σ^2 : $E[S^2] = E\left[\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}\right] = \frac{1}{n-1}E\sum_{i=1}^{n}(X_i^2 - 2X_i\bar{X} + \bar{X}^2)$

$$= \frac{1}{n-1}E\left[\sum X_i^2 - 2\bar{X}\sum X_i + \bar{X}^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$$

Recall
$$E[\bar{X}^2] = var(\bar{X}) + (E[\bar{X}])^2 = \frac{\sigma^2}{n} + \mu^2$$

(using $\sum_{i=1}^{n} X_i = n\bar{X}$)

$$= \frac{1}{n-1}\sum E[X_i^2] - nE[\bar{X}^2]$$

$$= \frac{1}{n-1}\sum(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n}+\mu^2\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 + \cancel{n\mu^2} - \sigma^2 - \cancel{n\mu^2}\right) = \frac{1}{n-1}(n\sigma^2 - \sigma^2) = \sigma^2$$

(1.2) Mean Squared Error: The MSE of an estimator $T(\mathbf{X})$ of θ is

$$MSE(T(\mathbf{X})) = E[(T(x)-\theta)^2]$$

"Easier" interpretation:

$$MSE(T(X)) = E[T^2] - 2\theta E[T] + \theta^2$$

$$= \underbrace{E[T^2] - (E[T])^2}_{} + \underbrace{(E[T])^2 - 2\theta E[T] + \theta^2}_{}$$

$$= Var(T) + (E[T]-\theta)^2$$

And we define $Bias(T(\mathbf{X})) \equiv E[T(\mathbf{X})] - \theta$

Hence $MSE(T(X)) = var(T(X)) + Bias^2(T(X))$

$\Rightarrow$ MSE combines bias and variance of an estimator

(1.3) Next, we're gonna see two different methods of finding estimators (in Regression, we use Least Squares Est, but let's see 2 different ones):

(1.3.1) Method of Moments Estimator

Recall (from last lecture), the K^{th} Moment of a RV X is

$$E[X^k] = \begin{cases} \sum_x x^k f(x) & \text{if } X \text{ is discrete} \\ \int_{\mathbb{R}} x^k f(x)\,dx & \text{if } X \text{ is continuous} \end{cases}$$

discrete cts

Now suppose $X_1, \ldots, X_n \overset{iid}{\sim}$ some pmf/pdf $f(x)$.

The MoM estimator for $E[X^k]$ is $\sum_{i=1}^n X_i^k/n$

(Does this make sense? Absolutely. Recall that the Law of Large Numbers implies that $\sum_{i=1}^n X_i^k/n \to E[X^k]$ (as n gets large — LLN)

(Baby) Examples:

- MoM estimator for μ (1st Moment): $\mu = E[X_i^1]$ $\hat{\mu}_{MOM} = \sum_{i=1}^n \frac{X_i^1}{n} = \bar{X}$ //

- MoM estimator for σ^2 (variance is the second central moment, $E[(X-\mu)^2]$, but it can also be expressed as):

$$Var(X_i) = E[X_i^2] - (E[X_i])^2$$

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

obs: MoM of $E[X_i^2]$ (2ND Moment) $= \sum_{i=1}^n x_i^2/n$

Hence $\hat{\sigma}^2_{MOM} = \frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2 = \left(\frac{1}{n}\sum_{i=1}^n x_i^2\right) - \bar{X}^2 = \frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{n} = \frac{n-1}{n} s^2$ //

(1.3.2) Maximum Likelihood Estimators (MLE)

Consider an iid random sample $X_1, , X_n$ where each X_i has pdf/pmf $f(x)$.
Suppose θ is some unknown parameter from X_i.
The likelihood function is $L(\theta) = \prod_{i=1}^{n} f(x_i)$

And the MLE of θ is the value of θ that maximizes $L(\theta)$ or in other words

$$\hat{\theta}^{MLE} = \underset{\theta}{\operatorname{argmax}}\ L(\theta)$$

→ In most common applications, optimizing the log-likelihood will make this problem a lot easier. Recall log is a one-to-one, monotonically increasing function.
Some properties (base a is replaced by "e" in MLE → natural log)

① $\log_a(x \cdot y) = \log_a x + \log_a y$ ② $\log_a\left(\frac{x}{y}\right) = \log_a x - \log_a y$

③ $\log_a x^m = m \log_a x$ ④ $\log_a \sqrt[n]{x^m} = \log_a x^{\frac{m}{n}} = \frac{m}{n} \log_a x$

$\log_e e^x = a$ $e^a = x$

Also, recall $\log_e 1 = 0$ and $\frac{d(\log x)}{dx} = \frac{1}{x}$

Example: $X_1, \dots, X_n \overset{iid}{\sim} Bern(p)$. Find $\hat{p}_{MLE}$

Recall pmf of Bernoulli(p) is $f(x) = \begin{cases} p^x(1-p)^{1-x} & \text{if } x=0 \text{ or } x=1 \\ 0 & \text{otherwise} \end{cases}$

The likelihood: $L(p) = \prod_{i=1}^{n} f(x_i) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}$

The log likelihood: $\ell(p) = \sum_{i=1}^{n} x_i \log p + (n - \sum_{i=1}^{n} x_i) \log(1-p)$

Take partial derivative with respect to p and set to 0.

$$\frac{\partial \ell(p)}{\partial p} = \frac{\Sigma x_i}{p} - \frac{n - \Sigma x_i}{1-p} \equiv 0$$

$$\Leftrightarrow \frac{\Sigma x_i}{p} = \frac{n - \Sigma x_i}{1-p} \Rightarrow \Sigma x_i - \cancel{p\Sigma x_i} - pn + \cancel{p\Sigma x_i} = 0$$

$$\hat{p}_{MLE} = \frac{\Sigma x_i}{n} = \bar{X} \,//$$

(2) Sampling Distribution: statistics (recall some examples $\bar{X}$ & s^2) are RV's. So it's sometimes useful to figure out their distributions, which are called "sampling distribution".

Ex: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \sigma^2/n)$

(2.1) Chi-squared distribution (χ^2)

Def: if $Z_1, \ldots, Z_K \overset{iid}{\sim} N(0,1)$ then $Y = \sum_{i=1}^{K} Z_i^2$ has a χ^2_K dist

$\therefore$ $Y \sim \chi^2(K)$ (chi-squared dist. with K degrees of freedom)

and pdf is $f_Y(y) = \dfrac{1}{2^{K/2}\underbrace{\Gamma(K/2)}_{\text{gamma function}}} y^{\frac{K}{2}-1} e^{-y/2}$, for $y > 0$

$E[Y] = K$, $var(Y) = 2K$

Property: χ^2's add up. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \chi^2(d_i)\ \forall i$

then $\sum_{i=1}^{n} Y_i \sim \chi^2(\sum_{i=1}^{n} d_i)$

Usually χ^2 dist comes up when we try to estimate σ^2, for example, if $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \rightarrow s^2 = \dfrac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \sim \dfrac{\sigma^2}{n-1}\chi^2(n-1)$

(2.2) t distribution

Def: suppose that $Z \sim N(0,1)$, $Y \sim \chi^2(K)$ and Z and Y are indep.

$T \equiv Z / \sqrt{Y/K}$ has the student t dist. with K degrees of freedom

$\therefore$ $T \sim t(K)$ with pdf $f_T(x) = \dfrac{\Gamma\left(\frac{K+1}{2}\right)}{\sqrt{\pi K}\ \Gamma\left(\frac{K}{2}\right)} \left(\dfrac{x^2}{K} + 1\right)^{-\frac{K+1}{2}}$, $x \in \mathbb{R}$

- $K=1$ gives Cauchy dist, with really fat tails
- $t(K) \rightarrow N(0,1)$ as K becomes large

$E[T] = 0$ and $var(T) = \dfrac{K}{K-2}$ $(K > 2)$

Usually the t dist. is used for finding confidence intervals and conduct hypothesis tests for the mean μ.

2.3 F distribution

Def: Suppose that $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ and X, Y are indep.
Then $F \equiv \frac{mX}{nY}$ has the F dist with n and m degrees of freedom

$\therefore F \sim F(n,m)$ with pdf $f_F(x) = \dfrac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \dfrac{\left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1}}{\left(\frac{n}{m}x+1\right)^{\frac{n+m}{2}}}$, $x > 0$

F distribution is used to find confidence intervals and conduct hypothesis tests for the ratio of variances for 2 ≠ processes.
(Remember in Regression when we analyze the ANOVA Table, we get F-stats for testing the equality of treatment means, i.e.
$F = \frac{MSTreatment}{MSError}$, which comes from $\left\{ \begin{array}{l} E(MSTr) = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ E(MSE) = \sigma^2 \end{array} \right)$

$\hookrightarrow H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$

2.4 Normal (or Gaussian) distribution

Def $X \sim N(\mu, \sigma^2)$ if it has pdf $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, $x \in \mathbb{R}$

$E[X] = \mu$ and $var(X) = \sigma^2$

Thm Additive property of normals: If $X_1, \dots, X_n \overset{iid}{\sim} N(\mu_i, \sigma_i^2)$, $i = 1, \dots, n$
then $Y \equiv \sum_{i=1}^{n} a_i X_i + b \sim N\left(\sum_{i=1}^{n} a_i\mu_i + b, \sum_{i=1}^{n} a_i^2 \sigma_i^2\right)$
(i.e. a linear combination of normals is itself normal)

$\Rightarrow$ Fact: a normal dist. is completely characterized by its mean and variance (if we add up normals, the result is still normal, just need to compute new mean and variance).

Ex: $X \sim N(3,4)$ and $Y \sim N(4,6)$, X, Y are indep. Find dist of $2X - 3Y$
$E[2X-3Y] = 2E[X] - 3E[Y] = -6$ $var(2X-3Y) = 4v(X) + 9v(Y) = 70$
$\Rightarrow 2X - 3Y \sim N(-6, 70)$

2.4.1 Standard Normal

Def: The Normal (0,1) distribution is called standard normal, often denoted by Z (tables available for cdf).

Any normal can be standardized by applying the transformation $Z = (X-\mu)/\sigma$

pdf of N(0,1) (usually represented by ϕ)

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad z \in \mathbb{R}$$

cdf of N(0,1) (usually represented by Φ)

$$\Phi(z) = \int_{-\infty}^{z} \phi(t)\,dt, \quad z \in \mathbb{R}$$

Remarks

1. $P(Z \leq a) = \Phi(a)$
2. $P(Z \geq b) = 1 - \Phi(b)$
3. $P(a \leq Z \leq b) = \Phi(b) - \Phi(a)$
4. $\Phi(0) = 0.5$
5. $\Phi(-b) = P(Z \leq -b) = P(Z \geq b) = 1 - \Phi(b)$
6. $P(-b \leq Z \leq b) = \Phi(b) - \Phi(-b) = 2\Phi(b) - 1$

Then

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) = P(-k \leq Z \leq k) = \Phi(k) - \Phi(-k) = 2\Phi(k) - 1$$

$\hookrightarrow$ probability that any normal RV is within k std. dev of μ doesn't depend on μ or σ^2.

Ex: for $k=1$ standard deviation, $2\Phi(1) - 1 = 2(0.8413) - 1 = 0.6826$.
for $k=2$, $2\Phi(2) - 1 = 0.95$
for $k=3$, $2\Phi(3) - 1 = 0.997$

Example: $X \sim N(21, 4)$. Find $P(19 < X < 22.5)$

$$P(19 < X < 22.5) = P\left(\frac{19-21}{\sqrt{4}} < Z < \frac{22.5-21}{\sqrt{4}}\right)$$

$$= P(-1 < Z < 0.75)$$

$$= \Phi(0.75) - \Phi(-1)$$

$$= \Phi(0.75) - [1 - \Phi(1)]$$

$$= 0.7734 - [1 - 0.8413] = 0.6147$$

2.4.2 Some theorems / corollaries:

→ Corollary : $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N(\mu, \sigma^2/n)$
(dist of sample mean has normal dist and as the number of observations increases, $var(\bar{X})$ gets smaller).

→ Central Limit Theorem: suppose $X_1, \ldots, X_n$ iid with $E[X_i] = \mu$ and $var(X_i) = \sigma^2$. Then as $n \to \infty$

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

Remarks ① If n is large (usually $n \geq 30$ will do the job) → $\bar{X} \approx N(\mu, \sigma^2/n)$

② The X_i's don't need to be normal for the CLT to work. In fact, CLT works even on discrete distributions.

③ You can almost always use the CLT if the observations are iid.

Ex : $X_1, \ldots, X_{100} \overset{iid}{\sim} \exp(1/1000)$. Find $P(950 \leq \bar{X} \leq 1050)$

First : since X_i's are $\exp(\underbrace{1/1000}_{\lambda})$: $E[X_i] = \frac{1}{\lambda}$ $\quad var(X_i) = \frac{1}{\lambda^2}$

By CLT : $E[\bar{X}] = E[X_i] = 1/\lambda = 1000$

$Var(\bar{X}) = var(X_i)/n = 10000$

Hence $P(950 \leq \bar{X} \leq 1050) = P\left(\frac{950 - 1000}{\sqrt{10000}} \leq Z \leq \frac{1050 - 1000}{\sqrt{10000}}\right)$

$\approx P\left(-\frac{1}{2} \leq Z \leq \frac{1}{2}\right) = 2\Phi(0.5) - 1 = 0.383$

So the above probability can be approximated by Normal dist.
(The exact answer can be obtained by using Erlang distribution).

Summary

$$t_n = \frac{Z}{\sqrt{\frac{\chi^2_n}{n}}} \quad \text{indep}$$

$$F_{m,n} = \frac{\chi^2_m}{\chi^2_n} \quad \text{indep}$$

$$Z \sim N(0,1)$$

z_i's $\overset{iid}{\sim} N(0,1)$

$$\chi^2_n = \sum_{i=1}^{n} z_i^2$$

③ Confidence Intervals

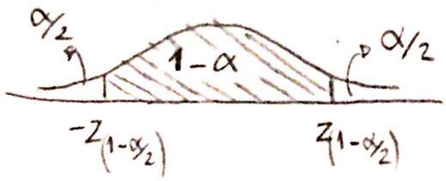
(3.1) Quantiles: the $(1-\alpha)$ quantile of a RV X is the value of x_α such that $P(X > x_\alpha) = 1 - F(x_\alpha) = \alpha$. Note that $x_\alpha = F^{-1}(1-\alpha)$ where $F^{-1}(\cdot)$ is the inverse cdf of X.

(3.2) Confidence interval: instead of estimating a parameter by a point estimator give a (random) interval that contains the unknown parameter with a certain probability ∴ there's a $(1-\alpha)$ chance that the parameter lies between the lower and upper limits.

- Two-sided CI: most times, we're interested in two-sided CI

⇒ $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, σ^2 known

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$



$$-Z_{1-\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq Z_{1-\alpha/2}$$

$(1-\alpha)$ CI is then:

$$\bar{X} \pm \underbrace{z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}_{\text{half-width}}$$

⇒ $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, σ^2 unknown
($n \leq 30$)

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \cdot \frac{\sigma}{s} = Z \cdot \frac{1}{\sqrt{\frac{s^2}{\sigma^2}}} = Z \cdot \frac{1}{\sqrt{\frac{(n-1)s^2}{\sigma^2} \frac{1}{n-1}}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} = t_{n-1}$$

$(1-\alpha)$ CI is

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

- One-sided CI (just interested in one bound

⇒ $100(1-\alpha)\%$ upper CI for μ

$$\mu \leq \bar{X} + z_\alpha \sqrt{\sigma^2/n}$$

⇒ $100(1-\alpha)\%$ lower CI for μ

$$\mu \geq \bar{X} - z_\alpha \sqrt{\sigma^2/n}$$

(use α quantile, not $\frac{\alpha}{2}$)