MATH 124 Spring 2005

Lecture: 17
Date: Apr 19, 2005

Skills you should acquire from this lecture:

- Approximations to sampling distribution for sample proportion and count random variables
- Finding mean and standard deviation for sample proportion.
- Sampling distribution for sample mean
- Central Limit Theorem (CLT)

Related readings in the textbook:

- Sections 5.1, 5.2
- Problems 5.55, 5.58
  (Discussed at TTh 8:10-9:25am)
- Problems 5.11, 5.15
  (Discussed at TTh 9:35-10:50am)

Sampling Distribution for Sample Proportion

For SRS situations with sample size n and population proportion p. The count of success X in the sample has binomial distribution $B(n, p)$

$B(n, p)$: n and p are parameters; B is notation for binomial.

Recall from our previous lecture that

$$\hat{p} = \frac{X}{n}$$

since X is a random variable with distribution $B(n, p)$ so is $\hat{p}$: a random variable. It can be shown that

$$\mu_{\hat{p}} = E[\hat{p}] = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

When the sample size is very large it can be shown that

X is approximately distributed $N(np, \sqrt{np(1-p)})$

and $\hat{p}$ is approximately distributed $N(p, \sqrt{\frac{p(1-p)}{n}})$

Reasonable rules of thumb for the approximation to be accurate

$$np \geq 10 \qquad n(1-p) \geq 10$$

Your textbook talks about a "continuity correction". You do not need to know about this.

What does the above mean? It means that we can use the normal distribution to get approximate probabilities for questions about binomial random variables and sample proportions.

eg suppose $n = 200$, $p = 0.4$ and X is Binomial(n, p)

What is probability

$$P(75 \leq X \leq 85)$$

since $np = 200(0.4) = 80$ $n(1-p) = 200(.6) = 120$

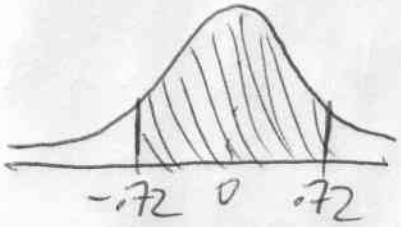
are both larger than 10 the approximation is reasonable.

$$\mu_X = np = 200(0.4) = 80$$

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{200(0.4)(0.6)} = 6.9282$$

So $P(75 \le X \le 85)$

$$= P\left(\frac{75-80}{6.9282} \le \frac{X-80}{6.9282} \le \frac{85-80}{6.9282}\right)$$



$$= P(-0.72 \le z \le 0.72)$$

$$= P(z < +.72) - P(z < -.72)$$

$$= .7642 - .2358$$

from table

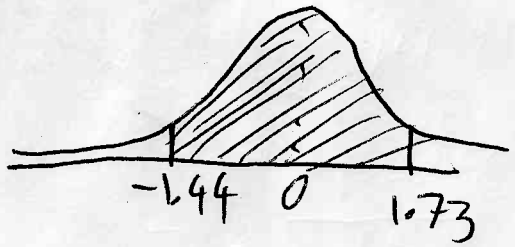
$$= .5284$$

What is probability $P(.35 \le \hat{p} \le .46)$

$$\mu_{\hat{p}} = .4 \qquad \sigma_{\hat{p}} = \sqrt{\frac{.4(.6)}{200}} = .0346$$

So $P(.35 \le \hat{p} \le .46) = P\left(\frac{.35-.4}{.0346} \le \frac{\hat{p}-.4}{.0346} \le \frac{.46-.4}{.0346}\right)$

$$= P(-1.44 \le z \le 1.73)$$



$$= P(z < 1.73) - P(z < -1.44)$$

$$= .9582 - .0749$$

$$= .8833$$

We have just discussed the sampling distribution for the sample proportion. In this class we concentrate most of our focus on two specific sample statistics, so next we move on to considering the sampling distribution of the sample mean

Sampling Distribution for the sample mean

Suppose that $X_1, X_2, \ldots, X_n$ are measurements of a r.v. X on n individuals chosen using a SRS. Note that X_i should therefore by independent

So sample mean

$$\bar{x} = \frac{\sum X_i}{n} = \frac{1}{n}\left[X_1 + X_2 + \ldots + X_n\right]$$

by the rules about expected (mean) values of r.vs

$$\mu_{\bar{x}} = \frac{1}{n}(\mu_X + \mu_X + \ldots \mu_X)$$

$$= \frac{n\mu_X}{n} = \mu_X$$

and the rules for variances

$$\sigma^2_{\bar{x}} = \left(\frac{1}{n}\right)^2 (\sigma^2_X + \sigma^2_X + \ldots + \sigma^2_X) = \left(\frac{1}{n}\right)^2 n\sigma_X^2 = \frac{\sigma_X^2}{n}$$

le

$\mu_{\bar{x}} = \mu_x$ (mean value of $\bar{x}$ is population mean of original data)

$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$ (std deviation of $\bar{x}$ is population std deviation of original data divided by squareroot of sample size)

Example

Suppose a chemistry measurement has mean 160 mg and standard deviation 1.5 mg

What is the mean and standard deviation of a single measurement?

$\mu_x = 160$ $\sigma_x = 1.5$

What is the mean and standard deviation of the mean of 3 measurements?

$\mu_{\bar{x}} = 160$ $\sigma_{\bar{x}} = \frac{1.5}{\sqrt{3}}$

What is the mean and standard deviation of the mean of 10 measurements

$\mu_{\bar{x}} = 160$ $\sigma_{\bar{x}} = \frac{1.5}{\sqrt{10}}$

What is mean and standard deviation of the sample mean of 25 measurements?

$\mu_{\bar{x}} = 160 \quad \sigma_{\bar{x}} = \frac{1.5}{\sqrt{25}} = .3$

Notice that the standard deviation of the sample mean goes down (ie gets smaller) as the sample size increases.

What is the distribution of $\bar{X}$?

If the original $X_1, \ldots X_n$ are all from a normal distribution with mean μ and standard deviation σ then $\bar{X}$ has exactly the normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$

Notation

$N(\mu, \sigma)$ is read as "Normal distribution with mean μ and standard deviation σ"

so For example

$N(\mu, \frac{\sigma}{\sqrt{n}})$ would be read as "Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$"

What if the distribution of $X_1, \ldots, X_n$ is not normal?

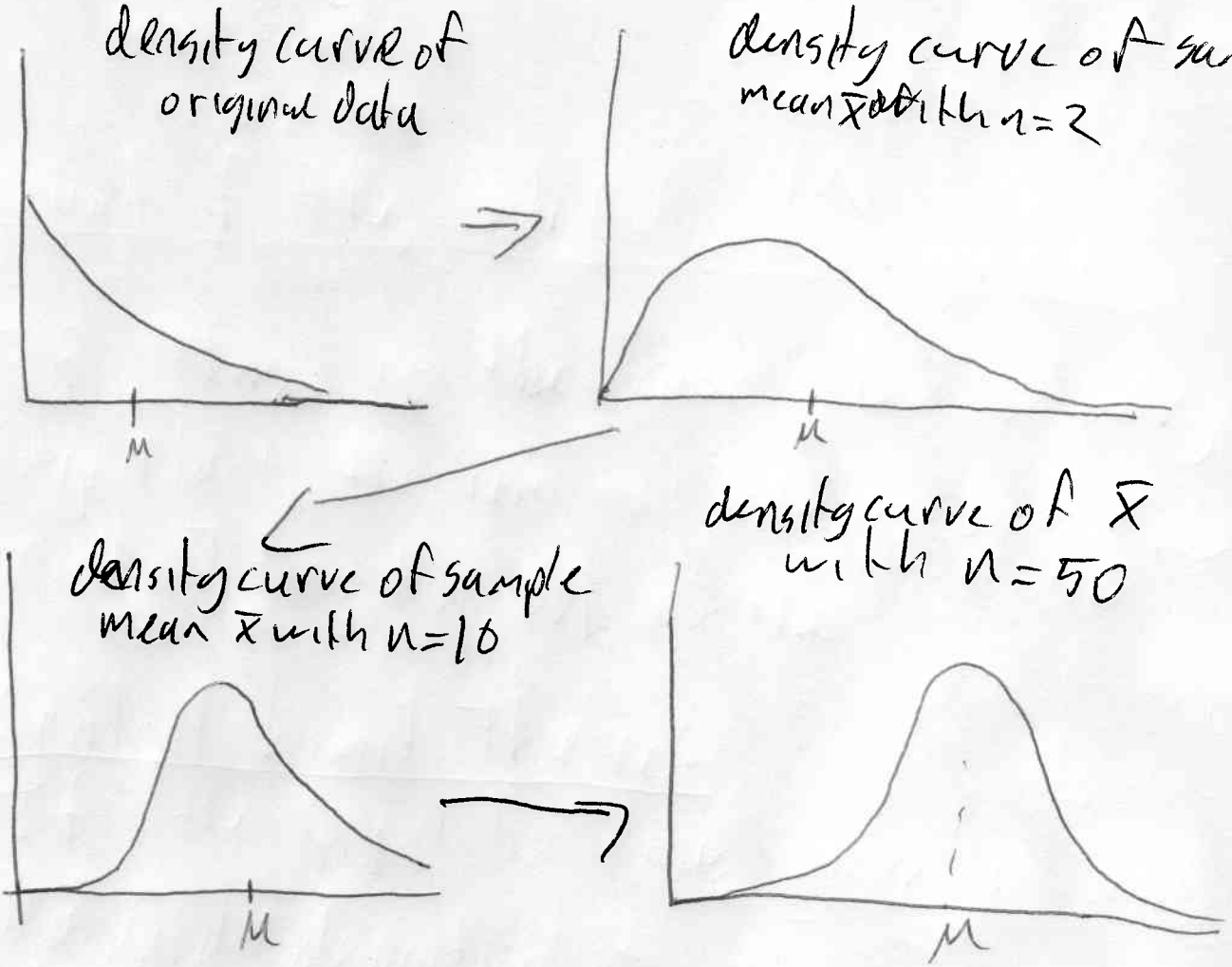
If this is the case then a very important theorem in statistics helps us out. It is known as the.....

Central Limit Theorem

If the population has mean μ and standard deviation σ, then no matter the distribution when we take a large sample the sampling distribution of $\bar{X}$ is approximately $N(\mu, \frac{\sigma}{\sqrt{n}})$

In otherwords when the sample size is large enough we can use the normal distribution when dealing problems about $\bar{x}$.

Graphical representation of CLT



ie the density curve of $\bar{x}$ approaches the bell shaped normal distribution as n gets larger.

Solutions For MWF 11-12

Problem 5.55

"The number of high school dropouts in 25,000"
$\equiv X$
is a binomial random variable with $n = 25000$
$p = .121$

(a) For a binomial random variable

mean $\mu_X = np$

std dev $\sigma_X = \sqrt{np(1-p)}$

so mean $\mu_X = (25000)(.121) = 3025$

and standard deviation $\sigma_X = \sqrt{25000(.121)(1-.121)}$
$= 51.5653$ (4dp)

(b) The probability statement is
$P(X \geq 3500)$

Because the sample size is large
X is approximately Normal with mean $\mu = 3025$
and standard deviation $\sigma = 51.5653$

This is the standardization step

So

$$P(X \geq 3500) \approx P\left(\frac{X-3025}{51.5653} \geq \frac{3500-3025}{51.5653}\right)$$

this is because of the normal approximation

$$= P(Z \geq 9.21)$$

$$= 1 - P(Z < 9.21)$$

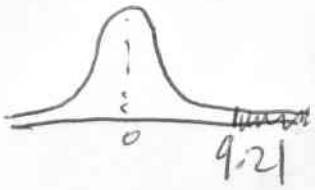
This symbol means "much much less"

From the table we know that

$$P(Z < 3.49) \ll P(Z < 9.21)$$

$$\Rightarrow .9998 \ll P(Z \leq 9.21)$$

$$\Rightarrow 1 - P(Z < 9.21) \ll .0002$$



Problem 5.58

$X \equiv$ "The number of red blossoms in n plants"

vs binomial randomvariable with $n=n$ $p=.75$

(a) $n=8$

$$P(X=6) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$= \binom{8}{6} (.75)^6 (1-.75)^{8-2}$$

$$= \frac{8!}{6!2!} (.75)^6 (.25)^2 = .3115 \quad (4dp)$$

(b) $n = 80$

because X is a binomial r.v

$\mu_X = np = (80)(.75) = 60$

(c) Probability statement is

$P(X \geq 50)$

Because n is large and more exactly

$np = (80)(.75) = 60 > 10$

and

$n(1-p) = (80)(.25) = 20 > 10$

X is approximately normal with mean 60

and standard deviation $\sqrt{80(.75)(1-.75)} = 3.8730$

So

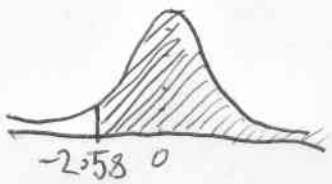
$P(X \geq 50) \approx P\left(\frac{X-60}{3.8730} \geq \frac{50-60}{3.8730}\right)$ ← Standardization step

This is because of the approximation

$= P(Z \geq -2.58)$

$= 1 - P(Z < -2.58)$

$= 1 - .0049 = .9951$



Solutions for MWF 2-3

Problem 5.15

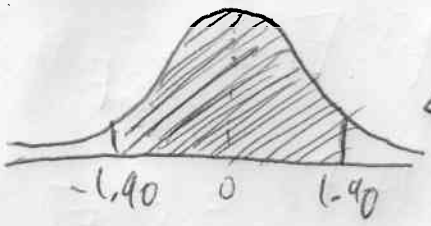
(a) When sample size is large sampling distribution of $\hat{p}$ is approximately normal with mean $\mu_{\hat{p}} = p = .51$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(.51)(.49)}{1005}}$

$= .0158$ (4dp)

Want $P(.48 \leq \hat{p} \leq .54)$ this is because of the approximation $\approx P\left(\frac{.48-.51}{.0158} \leq \frac{\hat{p}-.51}{.0158} \leq \frac{.54-.51}{.0158}\right)$

the standardization step

$= P(-1.90 \leq z \leq 1.90)$



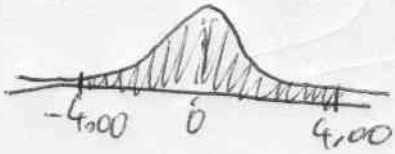
$= P(z \leq 1.90) - P(z \leq -1.90)$

$= .9713 - .0287$

$= .9426$

(b) $p = .06$ $\mu_{\hat{p}} = .06$ $\sigma_{\hat{p}} = \sqrt{\frac{.06(.94)}{1005}} = .0075$ notice this is smaller

want $P(.03 \leq \hat{p} \leq .09) \approx P\left(\frac{.03-.06}{.0075} \leq \frac{\hat{p}-.06}{.0075} \leq \frac{.09-.06}{.0075}\right)$



$= P(-4.00 < z < 4.00)$

≈ 1 (i.e. probability increased)

Problem 5.11

For McGwire $X \equiv$ "number of HR in 509 at bats"

X is binomial (approximately) with $n = 509$ $p = .116$

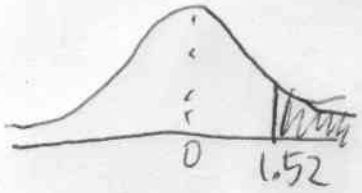
(a) $\mu_X = np = (509)(.116) = 59.044$

(b) $P(X \geq 70)$ ← probability statement

$$\sigma_X = \sqrt{np(1-p)} = \sqrt{509(.116)(1-.116)}$$
$$= 7.22$$

because n is large X is approximately normal with mean and sd above

$$P(X \geq 70) \approx P\left(\frac{X - 59.044}{7.22} \geq \frac{70 - 59.044}{7.22}\right)$$



$$= P(Z \geq 1.52)$$
$$= 1 - P(Z < 1.52)$$
$$= 1 - .9357$$
$$= .0643$$

(c) For Bonds $X \equiv$ "number of HR in 476 at bats"

X is binomial distributed with $n = 476$ $p = .0865$

$\mu_X = np = (476)(.0865) = 41.174$

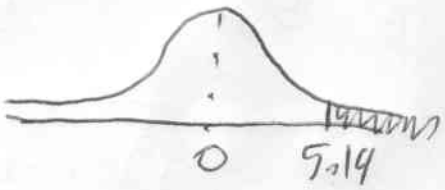
$\sigma_X = \sqrt{np(1-p)} = \sqrt{476(.0865)(1-.0865)}$

$= 6.1329$ (4dp)

X is approximately normal with mean 41.174 sd 6.1329

$P(X \geq 73) \approx P\left(\frac{X - 41.174}{6.1326} \geq \frac{73 - 41.174}{6.1326}\right)$

$= P(Z \geq 5.19)$



$= 1 - P(Z < 5.19)$

$\ll 1 - P(Z < 3.49)$

$= .0002$ (ie not very likely)

Based on these results it seems very unlikely given his past record that Barry Bonds would score 73 or greater home runs in a season just by chance. This means that for some reason, in his home run record season it is much more likely that something about his technique changed.