Lecture #12
6/15/11

Admin
→ HW posted - due Fri, come to office hrs if you have trouble
→ posted sol's & rubric

→ Also: JMP can be used on (1a) if you got q4 6 correct, otherwise do the pts manually that you missed

Midterms: added pts up to the median curve
Overall happy... breakdown p9
Ranges: $\geq 90 \Rightarrow A$, $\approx 32\%$
$70-89 \Rightarrow B$, $\approx 38\%$
$60-69 \Rightarrow C$, $\approx 19\%$
the rest in the tail
get it after class
"OK" $\Rightarrow$ means you followed through and get credit

Plan
- Admin
- Review of central theory
- CI's for prop's/philosophy
- CI's for mean data in general

We're going to do #5 right now since no one got it perfect
and the survey question

Yesterday we had the CLT for Bernoullis

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p),\ n \text{ large} \Rightarrow \hat{p} = \frac{X_1 + \ldots + X_n}{n} \sim N\left(p, \left(\sqrt{\frac{pq}{n}}\right)^2\right)$$

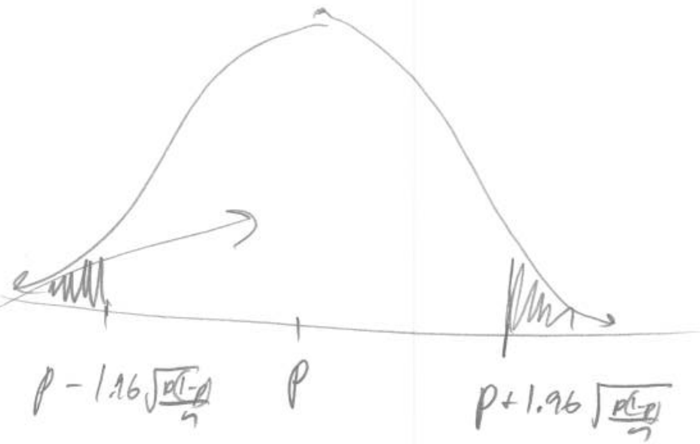
As well as the regular old CLT!

$$X_1, \ldots, X_n \overset{iid}{\sim} \text{something w/ mean } \mu, \text{ sd } \sigma,\ n \text{ large} \Rightarrow \bar{X} = \frac{X_1 + \ldots + X_n}{n} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

These two are the same but the top has more convenient notation because it reminds you that we care about proportion. We're going to use both of these today

Yesterday we talked about "control theory". We were asked to pick $\alpha = P(\text{type I error})$ and from there, we computed shutdown limits:

For $\alpha = 5\%$ (a standard choice), we get..



We made the probability statement

bracket notation indicates a set

$$P\left(\hat{p} \in \left[p \pm 1.96\sqrt{\frac{p(1-p)}{n}}\right]\right) = 95\%$$

pm is short for ...

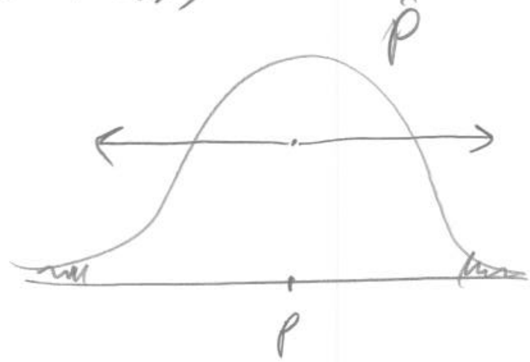
$[2,8] = [5 \pm 3]$

depends on your pt of view

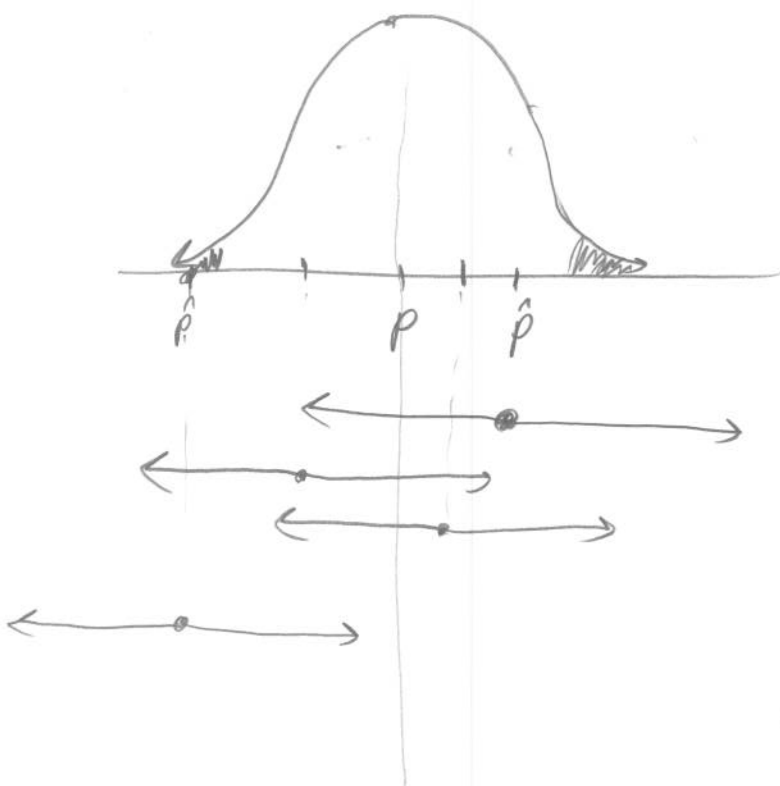
$$P\left(\hat{p} \in \left[p \pm z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right]\right) = 1-\alpha$$

remember this notation?

So we have a width that composes 95% of the mass



Now, imagine we start to do experiments. Take an SRS of n subjects, calculate little $\hat{p} = \frac{x_1+\dots+x_n}{n}$. Where can it land?



Construct the interval:

$$\left[\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}\right]$$

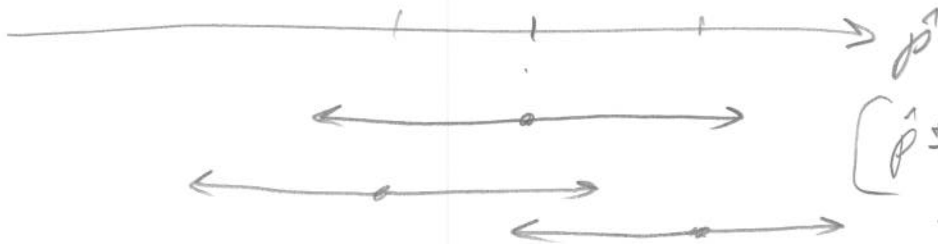
Now, draw a line down from p, the parameter

The one that fell outside the limits didn't "include" p, the real proportion.

How often does this happen? 5% of the time.

There are no secrets here. If we construct 20 such intervals we expect $E[S_{20}] = 20 \cdot 5\% = 1$ interval to be outside, but it's binomial, random!

Now, imagine the likely scenario, where we don't know p, the true mean. So erase this distribution



$$\left[\hat{p} \pm 1.96\sqrt{\frac{p(1-p)}{n}}\,\right]$$

here's the interval... what don't we know...?

Fix: $\left[\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\,\right]$ which is a close enough approx to the real thing

i.e. $SE[\hat{p}] = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ a nice luxury

$\Rightarrow$ $\hat{p} \pm 1.96 \cdot$ is as good as what we had before..

An interval that "catches" p 95% of the time

This is called a confidence interval (CI) and is denoted:

$$CI_{p,95\%} = \left[\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

for which param? — "confidence level" or "coverage" — our estimated result — $Z_{0.025}$ — something else approx. $SE[\hat{p}]$

In general...

$$CI_{p,1-\alpha} = \left[\hat{p} \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

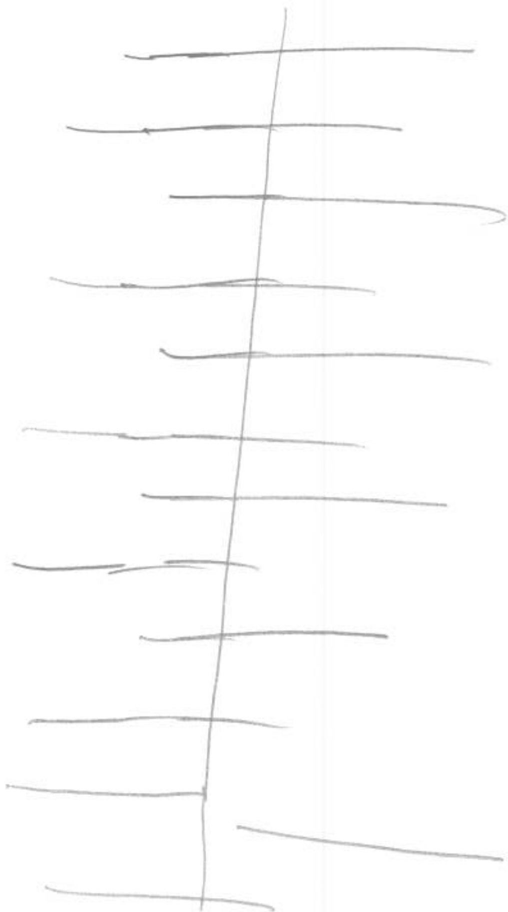
Let's do some CI's. We did a survey asking people after election if they voted for Obama. We asked 112 people, 58 said they did. Construct a 95% CI for the percent of people who vote for Obama.

$$\hat{p} = \frac{58}{112} = .518$$

$$CI_{p,95\%} = \left[\hat{p} \pm Z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = \left[.518 \pm 1.96\sqrt{\frac{.518 \cdot .482}{100}}\right] = [.42, .62]$$

Can we say anything about if Obama was going to win?

So again what is a 95% CI?



95% of experiments will capture the true p

~~It~~ It is not: $\cancel{P\left(p \in CI_{95\%}\right) = 95\%}$

wrong

Why? p is a #

this is like asking

$$P\left(\tfrac{1}{2} \in [0,1]\right) = 1$$

$$P\left(1.1 \in [0,1]\right) = 0$$

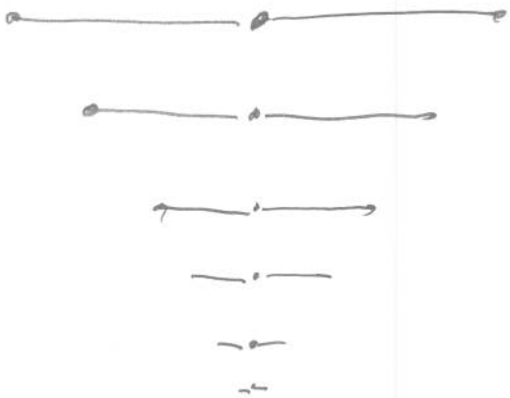
What does a CI ?

$$\left[\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

as n goes up...

width goes down

with the sq root of n



n=100

n=200

n=400

n=1,000

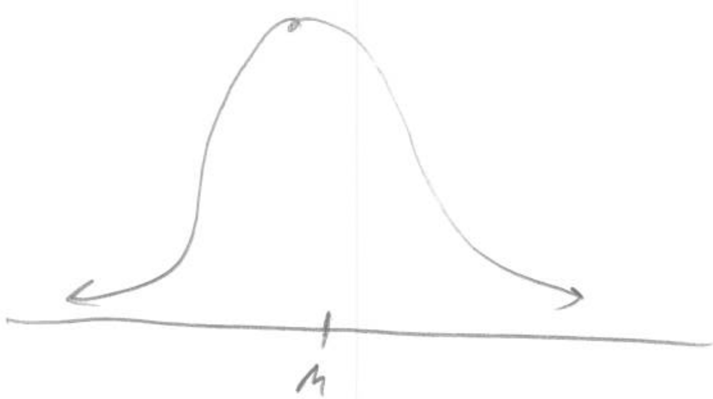
n=5,000

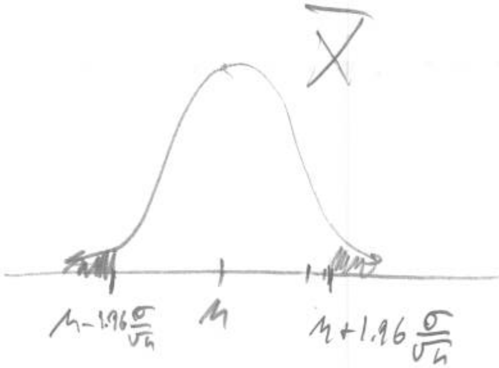
n=10,000

Now,

for samples.

$$\bar{X} \sim N\left(\mu, \left(\frac{\sigma}{\sqrt{n}}\right)^2\right)$$

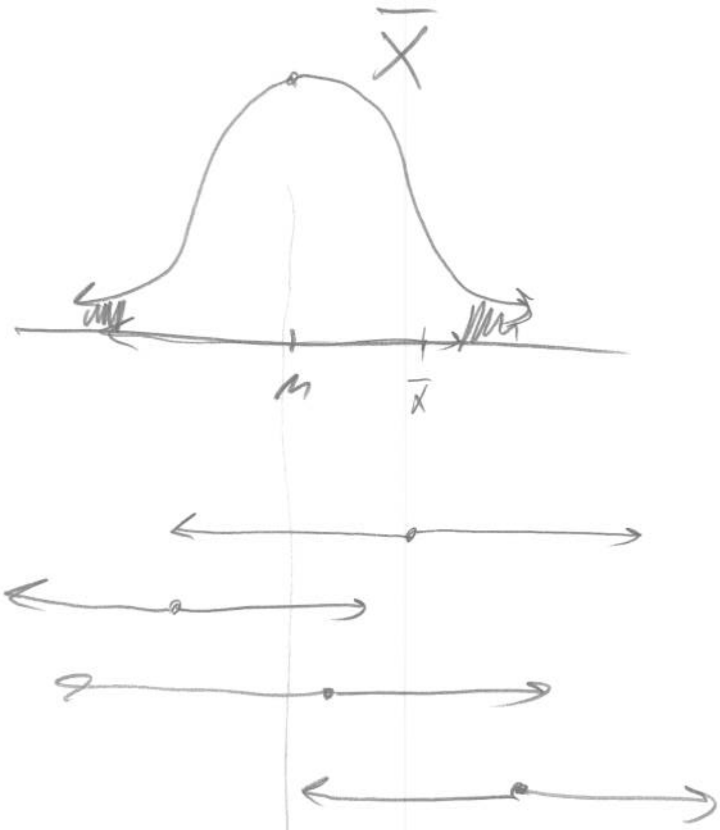




$$\Rightarrow P\left(\overline{x} \in \left[\mu \pm 1.96\frac{\sigma}{\sqrt{n}}\right]\right) = 95\%.$$

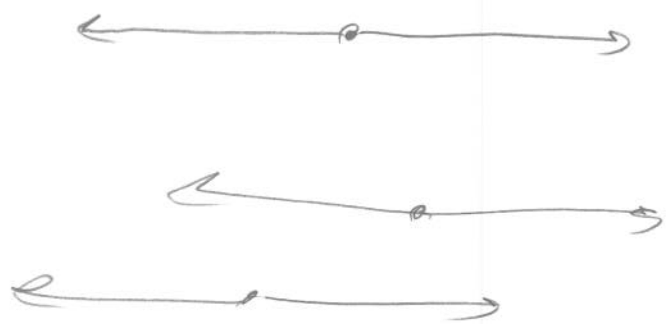
$$P\left(\overline{x} \in \left[\mu \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]\right) = 1 - \alpha$$

Now, draw a bunch of cap[illegible].



didn't make it

Once again, ~~the~~ the usual situation is, we don't know μ! But assume for now, we know σ



$$\left[\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}\right]$$

95% CI

In general $$\left[\bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Let's say you sample 70 people in ~~our class~~ UPenn, whose avg height is 70", we know $\sigma = 2$. Make a CI for height at UPenn

$$CI_{\mu, 95\%} = \left[70 \pm 1.96 \cdot \frac{2}{\sqrt{70}}\right] = [69.56, 70.44]$$