§12.2 Increasing accuracy

In parameter estimation, more accuracy means smaller standard error.
As we once saw in Ch 9, usually the only thing we have control over is n.
But in practice, n can't be infinite and maybe we don't even control it either.
What can we do?

ex: Let $\hat{\theta} = \bar{X}_1 - \bar{X}_2$ difference in means

For large sample $V(\hat{\theta}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$ (large sample)

and standard error is $\sigma_{\hat{\theta}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

Assume you can sample n times.
How do we choose n_1, n_2?
Use $n_1 + n_2 = n$ or $n_2 = n - n_1$.

The variance as a fcn of n_1 is

$$V_{\hat{\theta}}(n_1) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n - n_1}$$

We want to minimize V.

$$\frac{dV}{dn_1} = V'(n_1) = \frac{-\sigma_1^2}{n_1^2} - \frac{\sigma_0^2}{(n-n_1)^2}(-1)$$
$$= -\frac{\sigma_1^2}{n_1^2} + \frac{\sigma_0^2}{(n-n_1)^2}.$$

Note $V''(n_1) = \frac{2\sigma_1^2}{n_1^3} + \frac{2\sigma_0^2}{(n-n_1)^3} > 0.$

So any critical number is a local min.

$$V'(n_1) = 0 \Rightarrow -\sigma_1^2(n-n_1)^2 + \sigma_0^2 n_1^2 = 0$$

$$-\sigma_1^2 n^2 + 2\sigma_1^2 n n_1 - \sigma_1^2 n_1^2 + \sigma_0^2 n_1^2 = 0$$

$$(\sigma_0^2 - \sigma_1^2) n_1^2 + 2n\sigma_1^2 n_1 - n^2\sigma_1^2 = 0.$$

discriminant $(2n\sigma_1^2)^2 - 4(\sigma_0^2 - \sigma_1^2)(-n^2\sigma_1^2)$
$= 4n^2\sigma_1^4 + 4n^2\sigma_1^2\sigma_0^2 - 4n^2\sigma_1^4$
$= 4n^2\sigma_1^2\sigma_0^2.$

and $n_1 = \frac{-2n\sigma_1^2 \pm 2n\sigma_1\sigma_0}{2(\sigma_0^2 - \sigma_1^2)} = \frac{n\sigma_1 \cdot (\sigma_1 \pm \sigma_0)}{(\sigma_0^2 - \sigma_1^2)}$

WLOG, let $\sigma_0^2 > \sigma_1^2$. Then we require $\sigma_1 + \sigma_0$ term in top to have $n_1 > 0$.

$$n_1 = \frac{+n\sigma_1(\sigma_0 - \sigma_1)}{\sigma_0^2 - \sigma_1^2} = \left(\frac{\sigma_1}{\sigma_0 + \sigma_1}\right) n.$$

Note: This implies that to minimize V for difference in means we ~~sample~~ more of the type that has inherently large variance

$$n_1 \approx \frac{\sigma_1}{\sigma_2+\sigma_1} n \quad , \quad n_2 \approx \frac{\sigma_2}{\sigma_2+\sigma_1} n$$

ex: best-fit line $y = \beta_0 + \beta_1 x + \epsilon$.

Recall a C.I for β_i is $\hat{\beta}_i \pm t^* \sqrt{V(\hat{\beta}_i)}$

$$\Rightarrow \hat{\beta}_i \pm t^* c_{ii} \sigma^2$$

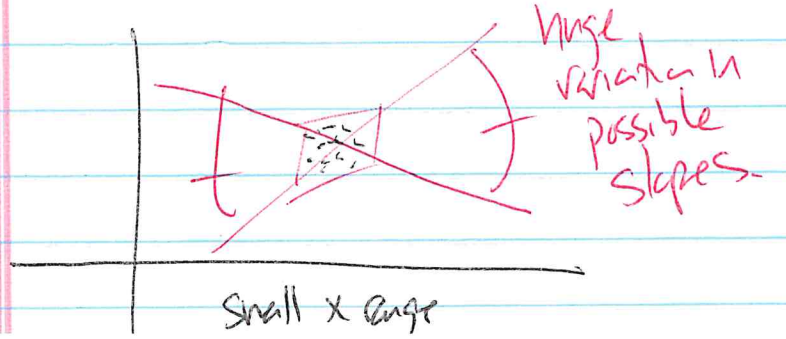
where c_{ii} is for $(X^T X)^{-1} = \begin{bmatrix} \Sigma x_i^2 / n S_{xx} & -\bar{x}/S_{xx} \\ -\bar{x}/S_{xx} & 1/S_{xx} \end{bmatrix}$

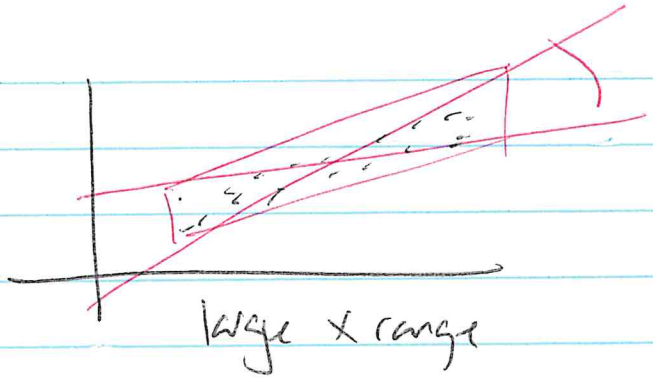
For estimating β_1, $V(\hat{\beta}_1) = c_{11}\sigma^2 = \frac{\sigma^2}{S_{xx}}$

where $S_{xx} = \sum (x_i - \bar{x})^2$.

* The more spread in x, the better our estimator for slope is.

Which kind makes sense...

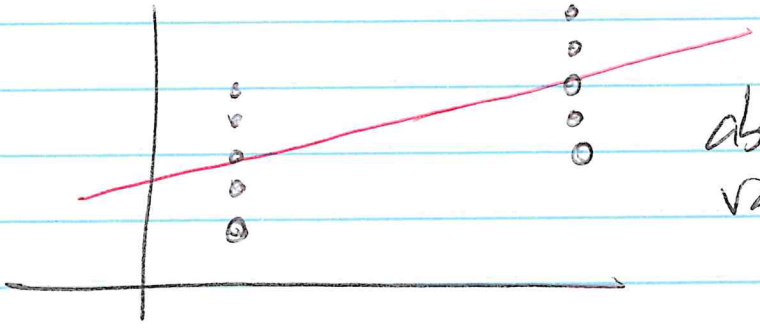




smaller possible range of slope.

large x range

But also feels weird... something like



absolutely minimizes variance.

§ 10.3 Matched-Pairs Experiment

Real-world experiments often measure a r.v. X pre- (X_1) and post- (X_2) treatment

As these measurements are on the same item or individual, clearly don't expect independence.

So, measuring effectiveness, we can't simply look at $\bar{X}_1 - \bar{X}_2$.

We use X_{1i} and X_{2i} to indicated the pre- and post- data point of the ith individual.

We use the standard assumption that X's are normally distributed: $X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$

No reason to expect either $\mu_1 = \mu_2$ or $\sigma_1^2 = \sigma_2^2$.

So $E(X_{1i}) = \mu_1$, $V(X_{1i}) = \sigma_1^2$
$E(X_{2i}) = \mu_2$, $V(X_{2i}) = \sigma_2^2$

but $Cov(X_{1i}, X_{2i}) = \rho \sigma_1 \sigma_2$

where $\rho = \dfrac{Cov(X_{1i}, X_{2i})}{\sigma_1 \sigma_2}$, the correlation coef from ch 5.

The stat we measure is the difference pre- to post-
$D_i = X_{1i} - X_{2i}$

Then $E(D_i) = \mu_1 - \mu_2$

$V(D_i) = V(X_{1i}) + V(X_{2i}) - 2Cov(X_{1i}, X_{2i})$

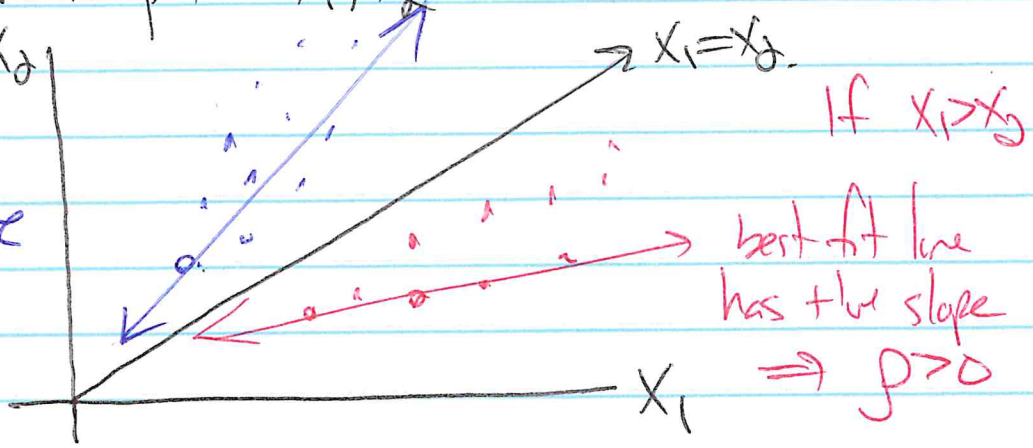
$= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$.

Then $\bar{D}_i \sim N(\mu_1 - \mu_2, V(D_i)/n)$, the sampling dist'n with $n-1$ df.

disc: why this is better.

① Commonly expect that a net treatment will have an effect that is relatively uniform.
either $X_1 > X_2$ or $X_1 < X_2$.

Consider the plot (X_1, X_2)



$X_1 < X_2$
again, slope is +ve
$\Rightarrow \rho > 0$

Thus $V(D_i) = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$

$< \sigma_1^2 + \sigma_2^2$ ← the variance if X_1, X_2 were independent.

The matched pairs stat yields a smaller confidence interval!

Summary: $D_i = X_{1i} - X_{2i}$ $\quad 1 \le i \le n$

We use t/2 methods to construct C.I. or hypo. tests.

Recall $n > 30$, using Z
$n \le 30$, using t.

Here $\bar{D} = \frac{\sum D_i}{n}$, $S_D^2 = \frac{1}{n-1}\sum_1^n (D_i - \bar{D})^2$.

$= \frac{1}{n-1}\left[\sum D_i^2 - n(\bar{D})^2\right]$

ex: testing effects of a supplement on cholesterol levels.

"randomly" select 6 people. (milligrams/deciliter)
X_1: C-level at beginning
X_2: C-level at 6 weeks

Really common assumption for biological things:
$X \sim N(\mu, \sigma^2)$

Person	1	2	3	4	5	6
X_1	210	235	208	190	172	244
X_2	190	170	210	188	173	228
D:	20	65	-2	2	-1	16

$\bar{D} = \frac{100}{6} = 16.7$

D^2: 400 4225 4 4 1 256

$S_{DD} = \sum D_i^2 = 4890$

$\sum S_D^2 = \frac{4890 - 6(16.7)^2}{5} = 642.332$

$S_D \approx 25.36$

Testing if change occurred $H_0: D = 0$
$H_a: D \neq 0$

$t = \frac{\bar{D} - D_0}{S_D/\sqrt{n}} = \frac{16.7 - 0}{25.4/\sqrt{6}} = 1.610$

p-value $= P(|t| > 1.610 \mid df = 5$
$= 2P(t > 1.610 \mid df = 5)$ > 0.10

Not significant.
Can not reject H_0.
No indication that the supplement caused a change in cholesterol.

(review)
Find a 90% C.I for μ_D.

$$\bar{D} \pm t_{0.05}(5) \, S_D/\sqrt{6}$$

$$16.7 \pm 2.015 \cdot \frac{25.36}{\sqrt{6}}$$

$16.7 \pm 2.015 \cdot 10.353$ (studied ever)

16.7 ± 20.854 or $-4.9 \leq \mu_D \leq 37.6$ *

Of course, we see $\mu_D = 0$ a possible value in the interval. This is exactly why we could not reject $H_0: \mu_D = 0$ in the prior hypo test

Is this really better?
The claim is, this C.I. is tighter than if we assumed independence.

Let's do $\mu_1 - \mu_2$ 40% C.I

Back to the data, we get

X_1: $n_1 = 6$, $\bar{X}_1 = 209.83$, $S_1^2 = 28.76$

X_2: $n_2 = 6$, $\bar{X}_2 = 193.16$, $S_2^2 = 495.36$

Using old diff in means methods,
note $df = n_1 + n_2 - 2 = 10$ is small.

we use

$$(\bar{X}_1 - \bar{X}_2) \pm t_{0.01}(10) S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}}$, pooled estimator.

Here $S_p^2 = \frac{S_1^2 + S_2^2}{2} = \frac{616.56}{2} = 24.71$

and standard error $S_p \sqrt{\frac{1}{6} + \frac{1}{6}} = 14.266$.

So $16.67 \pm 1.812(14.266)$

$\Rightarrow 16.67 \pm 25.85$ ← much bigger error term

↑ with basically the same center.

WORSE!