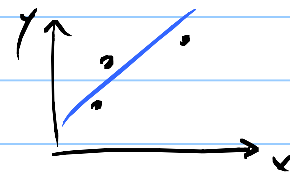
Lecture 14 (Ch 3-5)

overfitting in multiple regr.

We know That one can overfit data on x, y, if one uses a high-order polynomial in poly. regression. Recall That the main reason This can happen is because such a regression model will have a lot of parameters.

In multiple regression There is yet another way That overfitting can happen even w/o including high-order terms in The model:

Consider 3 cases on y and x_1:



A model like $y = \alpha + \beta x_1$ (a line)

cannot overfit That data

But a model like $y = \alpha + \beta_1 x_1 + \beta_2 x_2$ (a plain) overfits completely.

The reason is because in That case The 3 cases are in 3D (not 2D), and There is always one plain That goes Thru 3 points exactly.

Note That The additional variable x_2 can even be completely unrelated to y! It can even be just random values!

In other words, by arbitrarily making The space big, we opened up The possibility of overfitting.

So one can overfit even a multiple regression model without any non-linear (eg. quadratic, cubic, ...) terms.

You may Think This is happening only because I have 3 cases here. But even with more cases, one can still overfit by simply including more (even random) predictors in The model, if There are many more params in regression Than cases.

This overfitting problem is not specific to regression. ALL models can overfit when They are too large. CS students: WATCH OUT!

One last thing before we leave regression (until Ch. 11)

Here is an explanation of $df = n-1, n-2, \ldots, n-(k+1)$:

Q $\quad \bar{Y} = \frac{1}{n} \sum^{n} Y_i \qquad$ why n?

A $\quad \{Y_1, Y_2, \ldots Y_n\}$ are all indep. $\Longrightarrow$ $df = n$.

of numerator

Q $\quad s^2 = \frac{1}{n-1} \sum^{n} (Y_i - \bar{Y})^2 \qquad$ why $n-1$?

A $\quad \{Y_1 - \bar{Y}, Y_2 - \bar{Y}, \ldots, Y_n - \bar{Y}\}$ are not all indep.

There is 1 constraint on them: $\sum_i^n (Y_i - \bar{Y}) = 0$

$\therefore$ $df = n-1$.

of numerator

This is one reason why $\sum (Y_i - \bar{Y})^2$ is divided by $n-1$.

Similar reasoning implies that the df for SSE is $n-(k+1)$, which is why we define S_e^2 as $\frac{SSE}{n-(k+1)}$. Note for $k=1$ (i.e. simple linear regression), $S_e^2 = \frac{SSE}{n-2}$.

$\boxed{\text{This page is FYI}}$

In simple linear regression: $y = \alpha + \beta x$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{has} \quad df = n - 2 \leftarrow \text{2 constraints}$$

First constraint $\quad \sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$

Pf: $\frac{1}{n} \sum (y_i - \hat{y}_i) = \frac{1}{n} \sum (y_i - \hat{\alpha} - \hat{\beta} x_i)$

$$= \frac{1}{n} \sum y_i - \frac{1}{n} \sum (\hat{\alpha} + \hat{\beta} x_i)$$

$$= \frac{1}{n} \sum y_i - \left(\hat{\alpha} + \hat{\beta} \frac{1}{n} \sum x_i \right)$$

$$= \left[\bar{y} - \underbrace{(\hat{\alpha} + \hat{\beta} \bar{x})}_{= \bar{y} \text{ (2nd Normal equ.)}} \right] = 0$$

2nd constraint: $\quad \sum (y_i - \hat{y}_i) x_i = 0,$

Pf: $\frac{1}{n} \sum (y_i x_i - \hat{y} x_i) = \frac{1}{n} \sum [x_i y_i - x_i (\hat{\alpha} + \hat{\beta} x_i)]$

$$= \overline{xy} - \hat{\alpha} \bar{x} - \hat{\beta} \overline{x^2}$$

$$= \overline{xy} - \bar{x} (\bar{y} - \hat{\beta} \bar{x}) - \hat{\beta} \overline{x^2}$$

$$= (\overline{xy} - \bar{x}\bar{y}) - \hat{\beta} (S_x^2) = 0.$$

$$\hat{\beta} = \frac{\overline{xy} - \bar{x}\bar{y}}{S_x^2} \quad \text{(1st Normal equ.)}$$

C-table

This page is only FYI, for now

All of Ch.3 has been about understanding the relationship between several continuous variables. What about categ. vars?

For categorical data the relationship is best captured through the contingency table: C-table

aka confusion matrix.

Data

x	y
Yes	High
Yes	Low
Yes	High
No	High
Yes	High
No	Low
No	Low
perhaps	medium
perhaps	Low

	y: High	Low	Medium
x: Yes	3	1	0
No	1	2	0
perhaps	0	~~0~~ 1	1

∃ Relationship between x and y.

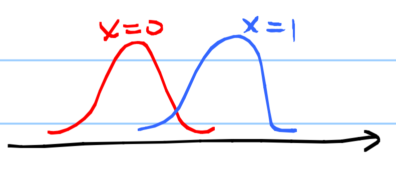
Maybe "positive" or "negative".

3 variables X, Y, Z ⟹ Cube = Set of Contingency Tables.

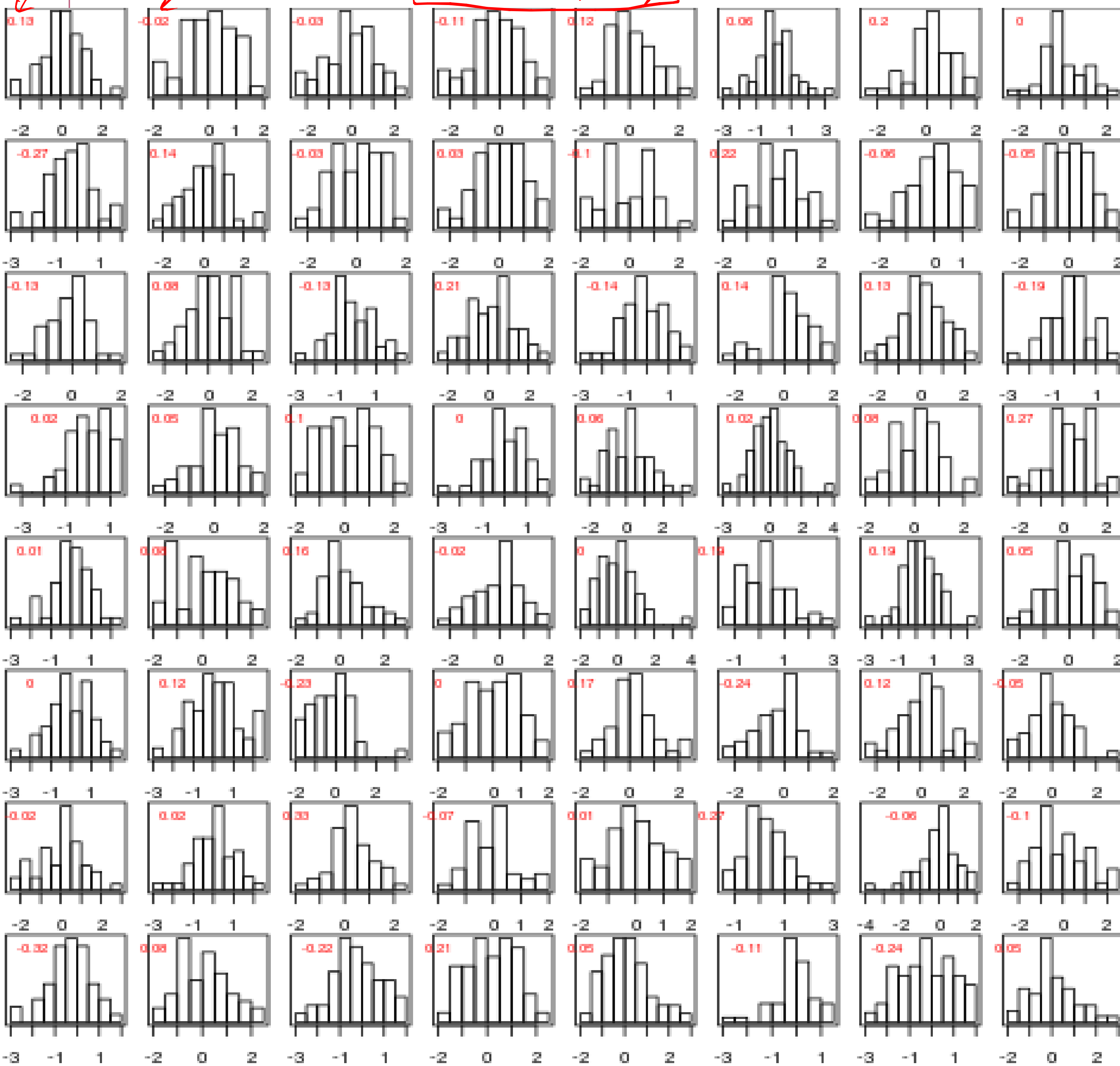
Q: What about mixed (discrete and cont)?

E.g. $\begin{cases} x = 0, 1 \\ y = \text{Continuous} \end{cases}$

A: Conditional histograms ⟹

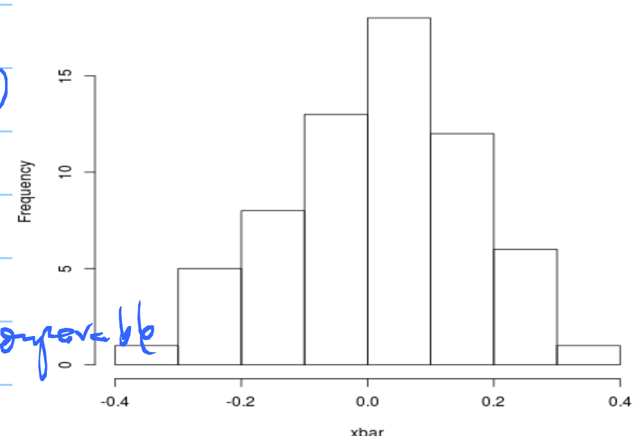


Ch 5 (5.5, 5.6)




```
ntrial = 64
xbar = numeric(ntrial)
par(mfrow=c(8,8))
 for( trial in 1:ntrial ){
 x = rnorm(50, 0, 1)
 hist(x, breaks=10)
 xbar[trial] = mean(x)
 }
hist(xbar, main="")
```

← Try rexp(50,1)



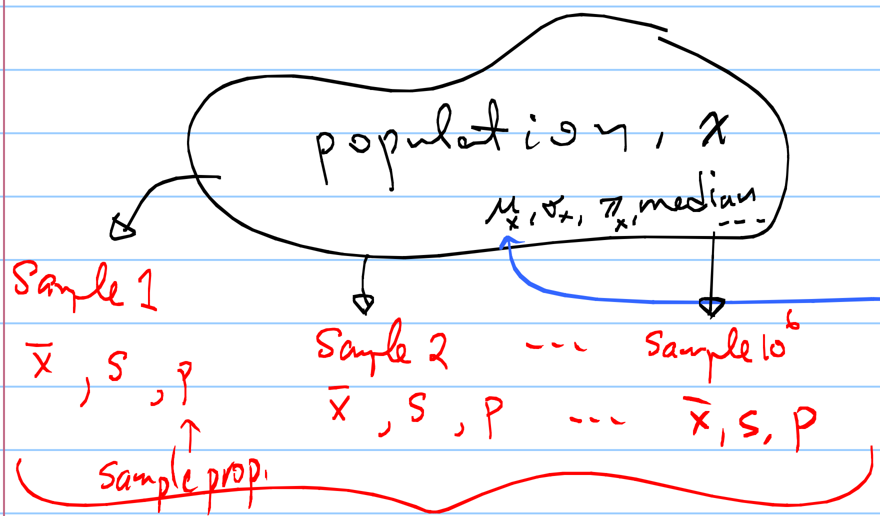
Q: What's $\bar{x}$ in each hist above?
What's the mean of the $\bar{x}$'s ?

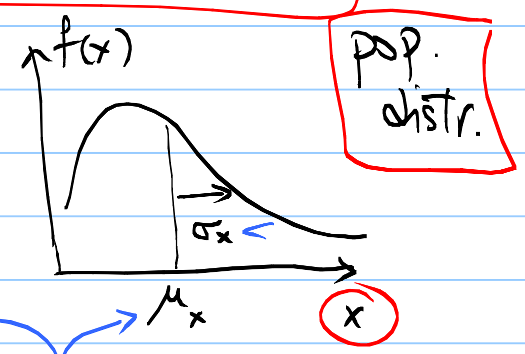
comparable

Q: What's s in each hist above?
What's s of the $\bar{x}$'s ?

very different!

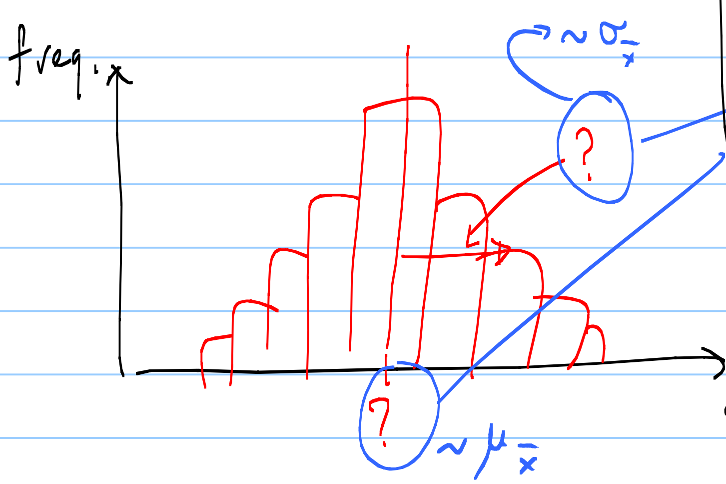
Sampling Distribution : Extremely Important !!





- pop./distr. mean
- " std. dev.
- " median,
- " proportion (of boys)

$\Rightarrow$ 10^6 $\bar{x}$'s $\Rightarrow$ histogram
or 10^6 s's



2 important quantities. One estimates the pop. parameter (e.g. μ), the other tells us how certain (precise) that estimate is.

Statistic ($\bar{x}$) or s or p } random vars.

The sampling dist. (of the sample mean) is a distribution, ie. a $p(x)$ or an $f(x)$ that can be derived mathematically, or simply assumed as a description of the population of all $\bar{x}$'s.
The only reason I talk about a histogram is to make the concept of the sampling dist. more intuitive. The histogram is sometimes called the "empirical sampling dist."

Note that the sampling distr. is the distribution of a sample statistic.
For example, the sampl. distr. of the sample mean,
tells us how the sample means are distributed.

Similarly, the sampl. distr. of the sample proportion,
tells us how the sample proportions are distributed. Etc.

Q What is the sampling distr. of $\bar{x}$? Normal, Poisson, ...?
A Later!

But even without knowing the dist., we can still find its
mean ($E[\bar{x}]$ or $\mu_{\bar{x}}$) and variance ($V[\bar{x}]$ or $\sigma_{\bar{x}}^2$):

If the population (ie. distribution) has mean μ_x and std. dev. σ_x, then

Mean of the sampling distr. of sample mean ($\mu_{\bar{x}}$):

Std. dev. " " " " " " " ($\sigma_{\bar{x}}$):

$$\mu_{\bar{x}} = E[\bar{x}] = \mu_x \quad \leftarrow \text{pop. mean}$$

$$\sigma_{\bar{x}} = \sqrt{V[\bar{x}]} = \sigma_x / \sqrt{n}$$

σ_x ← pop. std. dev.

n ← sample size

proof, below.

"Sometimes called "standard error of mean."

Derivation: Suppose we do <u>not</u> know the distr. of the population ($p(x)$, $f(x)$), but we <u>do know</u> its μ_x and σ_x

Of course, if you do know the pop. distr., then you can compute μ_x, σ_x as before:

$$E[x] \equiv \mu_x = \sum_x x\, p(x) \qquad \left(\text{or } \int x f(x)\,dx\right)$$

$$V[x] \equiv \sigma_x^2 = \sum_x (x-\mu_x)^2 p(x) \qquad \left(\text{or } \int \cdots dx\right)$$

Recall, $E[ax] = aE[x]$, $V[ax] = a^2 V[x]$, a = constant. Then

$$\mu_{\bar{x}} = E[\bar{x}] = E\left[\frac{1}{n}\sum_i^n x_i\right] = \frac{1}{n}\sum_i \underbrace{E[x_i]}_{\mu_x\ \forall i} = \frac{1}{\cancel{n}}\mu_x \left(\sum_i^{\cancel{n}} 1\right) = \mu_x .$$

$$\boxed{E[\bar{x}] \equiv \mu_{\bar{x}} = \mu_x}$$

The i^{th} obs. is a random value, there is nothing special about the i^{th} obs.

So, just drop the "i". Then $E[x_i] = E[x] = \sum_x x\, p(x) = \mu_x$.

Alternatively, work out $E[x_i]$ for each i, e.g. $i=1$

$E[x_1] = \sum_{x_1} x_1\, p(x_1) = \mu_x$, $E[x_2] = \mu_x$, etc.

$$\sigma_{\bar{x}}^2 = V[\bar{x}] = V\left[\frac{1}{n}\sum^n x_i\right] = \left(\frac{1}{n}\right)^2 \sum_i \underbrace{V[x_i]}_{\sigma_x^2}$$

The var. of each element in the pop. is the var. of the pop.

$$= \left(\frac{1}{n}\right)^2 \sigma_x^2 \sum_{i=1}^n 1 = \frac{\sigma_x^2}{n} \Longrightarrow \boxed{\sigma_{\bar{x}} = \sqrt{V[\bar{x}]} = \frac{\sigma_x}{\sqrt{n}}}$$

$\bar{x} = \frac{1}{n}\sum^n x_i \longrightarrow \mu_x \qquad s_{\bar{x}} =$

In Summary:

$\mu_{\bar{x}} \equiv E[\bar{x}] = \mu_x$ Tells us That we can use The sample mean (from The one sample of size n) to estimate The pop. mean μ_x with accuracy. ← see bottom of page.

$\sigma_{\bar{x}} \equiv \sqrt{V[\bar{x}]} = \frac{\sigma_x}{\sqrt{n}}$ Tells us That The typical deviation in $\bar{x}$ is $\frac{\sigma_x}{\sqrt{n}}$, and so it tells us how precise is our estimate of μ_x. ← certain.

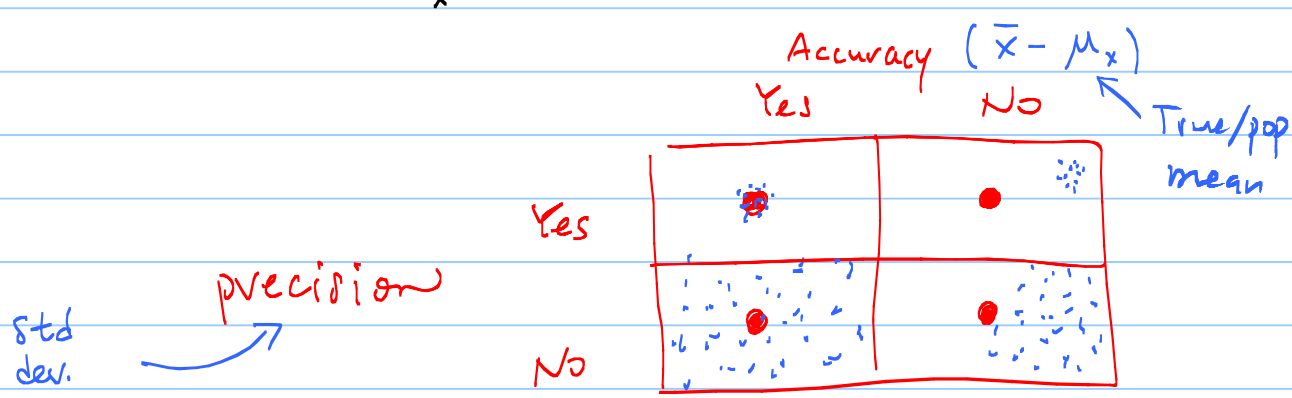
Note That $\mu_x, \sigma_x, \mu_{\bar{x}}, \sigma_{\bar{x}}$ are means and std. dev. of distributions, NOT of data. We are dealing with distributions, even though The thought exp. involved a hist.

$\mu_x = \sum_x x\, p(x)$, $\int x f(x)\, dx$; $\sigma_x^2 = \sum_x (x-\mu_x)^2 p(x)$, $\int (x-\mu_x)^2 f(x)\, dx$

FYI

$\underbrace{\bar{x} \text{ and } s_x}$ are measures of Accuracy & Precision;

and so $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$,



Now, what is the sampling distr. of sample means?

Thm If the pop. is Normal(μ, σ), then the sampling dist. of $\bar{x}$ is Normal with params: $N\left(\mu_{\bar{x}} = \mu_x = \mu,\ \sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}} = \frac{\sigma}{\sqrt{n}}\right)$

Central Limit Theorem (CLT)

even if the pop. is NOT normal, as long as $n =$ large (say > 30)

I'll go over this again tomorrow.

Now that we know the distr. of $\bar{x}$, we can compute probs. pertaining to a random (future) $\bar{x}$. eg. prob$(a < \bar{x} < b)$:

1a) $\boxed{\text{If}}$ pop. distr. $(p(x), f(x))$ is given, use it to compute μ_x, σ_x:

Eg. $\mu_x \equiv E[x] = \sum x\, p(x)$, $\sigma_x^2 \equiv V[x] = \sum (x-\mu_x)^2 p(x)$.

1b) $\boxed{\text{If}}$ pop. distr. is not known, assume its μ_x, σ_x (Ch. 7, 8)

2) CLT $\Rightarrow$ $\bar{x}$ is distributed as $N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right)$

3) Standardize: $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_x}{\sigma_x/\sqrt{n}} \sim N(0,1)$

4) prob$(a < \bar{x} < b)$ sample mean. Think about the meaning of this prob.

$$= \text{prob}\left(\frac{a-\mu_x}{\sigma_x/\sqrt{n}} < \underbrace{\frac{\bar{x}-\mu_x}{\sigma_x/\sqrt{n}}}_{z \sim N(0,1)} < \frac{b-\mu_x}{\sigma_x/\sqrt{n}}\right)$$

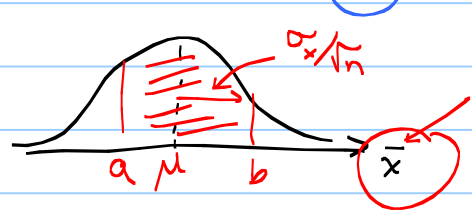


Table I.

hw-lect 14-1

Overfitting occurs in multiple regression even without higher powers of the predictors. Let's see it. Consider the data on y and x1 made here:

```
set.seed(123)
n = 10
x1 = runif(n,-1,1)
y = 1 + 2*x1 + rnorm(n,0,1)
```

a) Make the scatterplot of y vs x1.
b) Perform simple linear regression, and report the R^2.
c) Generate another n cases from runif(-1,1), and call that data x2. Then repeat this step three more times to generate x3, x4, and x5. In other words, in this step, generate data on x2,x3,x4,x5, where they are all independent of each other, and none of them are related to y.
d) Perform multiple linear regression on y,x1,x2,x3,x4,x5, and report R^2.

hw-lect 14-2 a) Write R code to produce The sampling distribution of The sample maximum, for samples of size 50 taken from a standard Normal. Use 5000 trials.
b) Then, repeat but for sample minimum.

Turn-in The code, and The resulting 2 histograms.

FYI, These distributions arise naturally when one tries to model extreme events, e.g. The biggest storms, The strongest earthquakes, The brightest stars, The smallest forms of life, etc.

hw-lect 14-3 Write R code to take 5000 samples of size n=100 from an exponential distr. with parameter $\lambda=2$, and plot a qqplot of The 5000 means. Recall That if The qq plot is a straight line, Then The histogram of The sample means is Normal. This will show That The sampl. dist. of sample means is Normal, even when The pop. is not!

hw_lect14-4

A sampling distribution (e.g. of the sample mean) is a distribution, not a histogram of observed sample means; the histogram of sample means discussed in class is just an intuitive way of thinking about the sampling distribution; technically, it's called the *empirical* sampling distribution. Of course, if the number of trials is infinite, then the empirical sampling distribution (i.e., the histogram) approaches the distribution. Anyway, to show that the sampling distribution is truly a distribution (not a histogram), let's derive one mathematically - no data at all.

Consider a population described by a Bernoulli random variable, i.e., x = 0,1, following the Bernoulli distribution, i.e., p(x) = pi^x (1-pi)^(1-x) . Suppose we take samples of size 2.
a) Write down all the possible samples. Hint: there are only 4.
b) For each of the possible samples, compute the sample mean.
c) For each of the possible samples, compute the probability. Hint: Use Bernoulli.
d) Based on your answers to parts a-c, find the probability of each of the possible sample means.

Note: your answer to part d *is* the sampling distribution of the sample mean! Note that it's not a histogram, but a real distribution.

This document was created with Win2PDF available at http://www.win2pdf.com.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.
This page will not be added after purchasing Win2PDF.