Lecture 15 (CR3, 5.5, 5.6, 7)

df

This page FYI

Why $(n-1)$, $(n-2)$, $n-(k+1)$, ... ?

Q $\bar{y} = \frac{1}{n}\sum_i^n y_i$ why n?

A $\{y_1, y_2, \dots, y_n\}$ are all independent $\Rightarrow$ $df\left(\text{of } \sum_i y_i\right) = n$

Q $s^2 = \frac{1}{n-1}\boxed{\sum_i^n (y_i - \bar{y})^2}^{\,S_{YY}}$ why $(n-1)$?

A $\{(y_1-\bar{y}), (y_2-\bar{y}), \dots, (y_n-\bar{y})\}$ are NOT all indep.

There is 1 constraint on them $\sum_i^n (y_i - \bar{y}) = 0$

I.e. There are $(n-1)$ indep. terms $\Rightarrow$ $df(\text{of } S_{YY}) = n-1$

[There are other reasons for $(n-1)$, too].

Q $s_e^2 = \frac{1}{n-2}\boxed{\sum_i^n (y_i - \hat{y}_i)^2}^{\,SSE}$ why $(n-2)$?

$\{y_i - \hat{y}_i\}$ satisfy 2 constraints (below) $\Rightarrow$ $df(\text{of } SSE) = n-2$

1^{st} constraint: $\underbrace{\frac{1}{n}\sum_i (y_i - \hat{y}_i)} = \frac{1}{n}\sum (y_i - \hat{\alpha} - \hat{\beta} x_i) = \bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ (see $\hat{\alpha}$ eqn.)

2^{nd} constraint: $\underbrace{\frac{1}{n}\sum_i (y_i - \hat{y}_i)x_i} = \frac{1}{n}\sum_i (y_i x_i - (\hat{\alpha} + \hat{\beta} x_i) x_i) = \overline{xy} - \hat{\alpha}\bar{x} - \hat{\beta}\,\overline{x^2}$

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

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

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

Ch 5 (5.5, 5.6) "Bridge between 1^{st} & 2^{nd} half of course"

$\boxed{\text{Sampling Distribution}}$: $\boxed{\text{Extremely Important!!}}$

population, x

μ_x, σ_x, π_x, median ---

The observed sample

Sample 1 of size n

$\bar{x}$, s, p

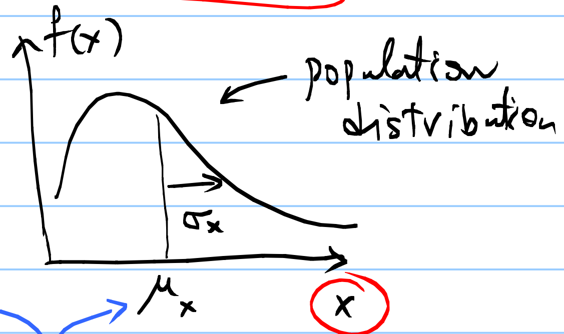
sample prop.

Sample 2 --- Sample 10^6

$\bar{x}$, s, p --- $\bar{x}$, s, p

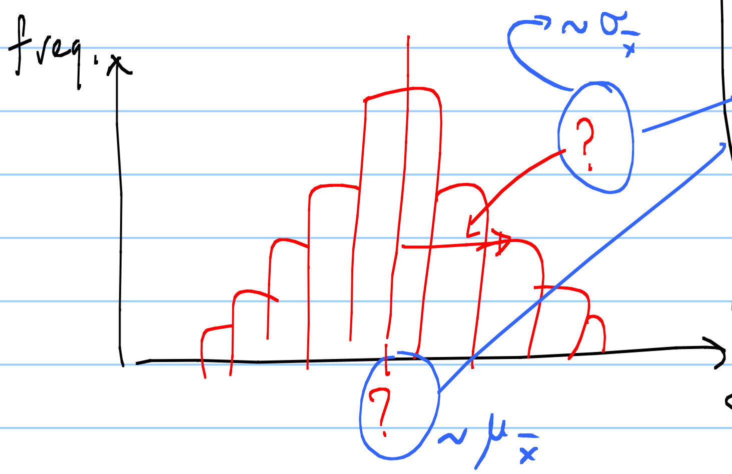
The imagined samples

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



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

or 10^6 s's



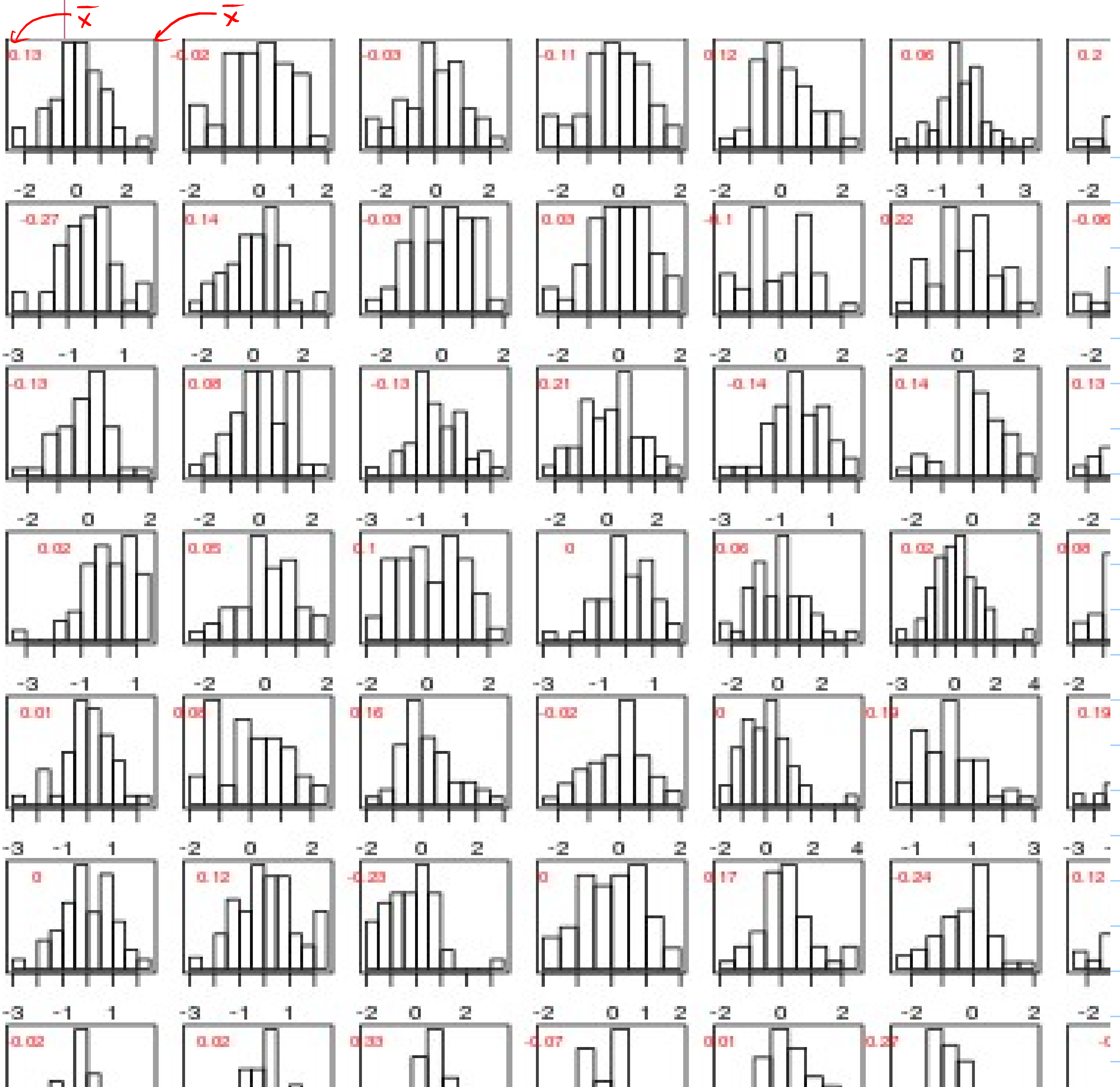
2 important quantities. One estimates the pop. parameter (e.g. μ), the other tells us how certain that estimate is.
precise

Statistic ($\bar{x}$, or s, or p, ---)

$\therefore$ $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$ of the sampling dist. tells us something about μ_x (the pop. mean!) and our uncertainty in it.

1) That's the sampling dist. of the sample mean. But one can talk about the sampling dustribution of the sample median, the sample standard deviation, the sample proportion, the sample anything.... Later (in ch 11), we will even talk about the sampling distribution of the sample fit. They all say something about the typical value and the typical fluctuation of the respective quantity.

2) Like the name suggests, it is a distribution, ie.p(x) or f(x), that can be derived mathematically, or simply assumed as a description of the population of all x's. In fact, you have already seen some sampling distributions, e.g. distribution of minimum, maximum, ... of sample of size 2 or 3 taken from Bernoulli.The only reason I talk about a histogram is to make the concept of the sampling dist. more intuitive; the talk of taking a zillion samples etc. is just thought experiment; in practice, we take only one sample of size n. The histogram is sometimes called the "empirical sampling dist."




```
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="")
```

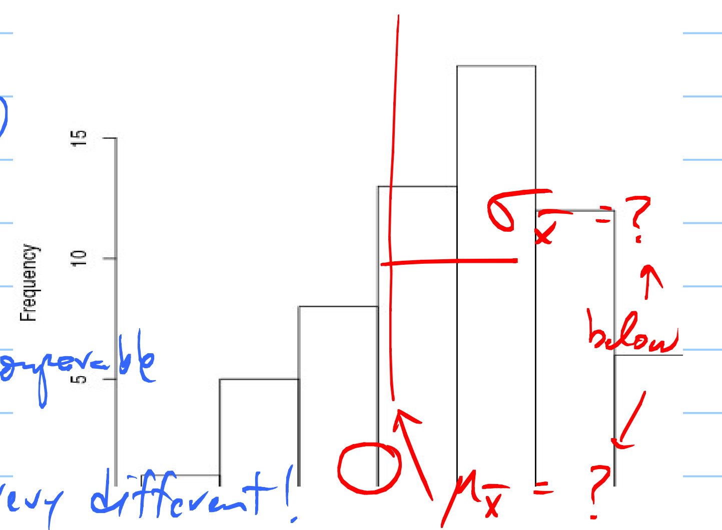
← Try rexp(50,1)

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!



Q What $\underline{\underline{is}}$ the sampling distr. of $\bar{x}$? Normal, Poisson, ...? 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 $\sigma^2_{\bar{x}}$):

Theorem: If The pop. (dist.) has mean & std. dev. μ_x, σ_x, Then

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

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

$\sigma_x \leftarrow$ pop. std. dev.

$n \leftarrow$ sample size

proof next page

sometimes called "standard error of mean."

where $\mu_{\bar{x}}$ = Mean of The Sampling distr. of Sample mean

$\sigma_{\bar{x}}$ = Std. dev. " " " " " " "

General notation/jargon:

statistics	pop. parameters
$\bar{x}$ (sample mean) is a point estimate of	μ_x (pop. mean)
s (" std. dev.)	σ_x (" std. dev.)
p (" prop.)	π_x (" prop.)

n (" size) is ~~NOT~~ related to pop. size. $\leftarrow$ for us $= \infty$

$\mu_{\bar{x}}$ = mean of The sampling dist. of $\bar{x}$

$\sigma_{\bar{x}}$ = std. dev. " " " " "

We skipped sec. 3-6, but it has one result That we need

constant

$$E[ax] = a\,E[x] \qquad V[ax] = a^2\,V[x]$$

$$E[x \pm y] = E[x] \pm E[y]$$

always +

$$V[x \pm y] = V[x] + V[y] + 0$$

x, y = indep.

partial proof below

Proof: Suppose we do not know the distr. of the population ($p(x)$, $f(x)$), but we do know 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:

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

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

Now, start!

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

$\Downarrow$

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

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.

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

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

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

(*) Although, $E[\,]$ and $V[\,]$ are mathematical operations that we defined in Ch. 2, e.g. $E[x] = \sum_x x\, p(x)$ and $V[x] = \sum_x (x-\mu_x)^2 p(x)$, it may help (intuitively) to think of $E[\,]$ and $V[\,]$ as mean and variance across the zillion imaginary samples taken from the pop.

FYI

Pf. of $E[aX] = aE[X]$:

$E[aX] = \int (ax) f(x)\, dx = a \int x f(x)\, dx = a E[X]$.

Pf. of $E[X+Y] = E[X] + E[Y]$:

With 2 variables (i.e. X, Y), E[ ] is defined this way

$$E[X] = \iint x\, f(x,y)\, dx\, dy.$$

Then $E[X+Y] = \iint (x+y)\, f(x,y)\, dx\, dy$

$$= \iint x\, f(x,y)\, dx\, dy + \iint y\, f(x,y)\, dx\, dy$$

$$= \int x \underbrace{\left(\int f(x,y)\, dy\right)}_{f(x)} dx + \int y \underbrace{\left(\int f(x,y)\, dx\right)}_{f(y)} dy$$

$$= \int x f(x)\, dx + \int y f(y)\, dy$$

$$= E[X] + E[Y].$$

The Pf. of $V[X \pm Y] = V[X] + V[Y] + 0$ is more complex.

FYI

Accuracy vs. precision

$\mu_{\bar{x}} \equiv E[\bar{x}] = \mu_x$ 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 box, below)

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

Note that $\mu_x, \sigma_x, \mu_{\bar{x}}, \sigma_{\bar{x}}$ are means and std. dev. of distributions, not of data, even though the thought-experiment led to a histogram. That's why we use μ, σ, E, V notation.

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

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

Accuracy $(\bar{x} - \mu_x)$ ← True/pop mean

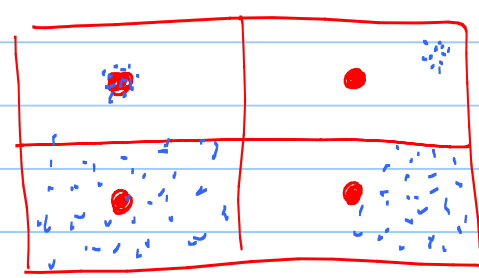
Yes No

Yes

No

precision

std dev.



Q: OK, so now we know $\mu_{\bar{x}} = \mu_x$ and $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$. But what is the distribution of $\bar{x}$ itself?

A: The Central Limit Theorem (CLT):

strong version: If $x \sim$ any dist. with mean $= \mu_x$, var. $= \sigma_x^2$
(if n = large)
not too important.

Then $\bar{x} \sim N(\underbrace{\mu_x}_{=\mu_{\bar{x}}}, \underbrace{\frac{\sigma_x}{\sqrt{n}}}_{=\sigma_{\bar{x}}})$

In English: For any pop. with mean μ_x and variance σ_x^2, the sampling dist. of the sample means is Normal with $\mu = \mu_{\bar{x}}$, $\sigma = \sigma_{\bar{x}}$, where we have already derived expressions for $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$, i.e.
$\mu_{\bar{x}} = \mu_x$, $\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$

Example: A sample of size 25 yields $\bar{x}_{obs} = 3$, $s_{obs} = 1.5$.
Suppose population is $N(\mu = 2, \sigma = 1)$. $\rightarrow \mu_x = 2$, $\sigma_x = 1$.
What's the prob of getting an even larger sample mean?

prob($\bar{x} > \bar{x}_{obs}$) =

$$\text{prob}(\bar{x} > 3) = \text{prob}(\overset{\frac{\bar{x} - \mu_x}{\sigma_x / \sqrt{n}}}{\overset{\parallel}{z}} > \frac{3-2}{1/\sqrt{25}}) = \text{prob}(z > 5) \approx 0 !$$

Just, in passing, note this:
This small prob suggests that $\mu = 2$ is a bad assumption.
In fact, the data suggest that μ is greater than 2 (closer to 3).
We will formalize these qualitative conclusions, below.

Important
Distinguish between random things (like x_bar $\bar{x}$) and non-random things (like x_bar_obs $\bar{x}_{obs}$ and mu_x μ_x). In lower-level stat classes this is not an important distinction; but at the 390 level, it is. And this important distinction will stay with us until the end of the quarter.

That was an example, and here is The general procedure for computing probs. of "Things" pertaining to The sampling dist, e.g. $pr(a<\bar{x}<b)$

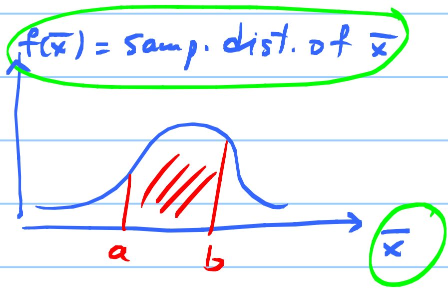
From CLT we know $\bar{x} \sim N(\mu_x, \frac{\sigma_x}{\sqrt{n}})$

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

Finally $pr(a<\bar{x}<b) = pr(a-\mu_x < \bar{x}-\mu_x < b-\mu_x)$

$$= pr\left(\frac{a-\mu_x}{\sigma_x/\sqrt{n}} < \frac{\bar{x}-\mu_x}{\sigma_x/\sqrt{n}} < \frac{b-\mu_x}{\sigma_x/\sqrt{n}}\right)$$

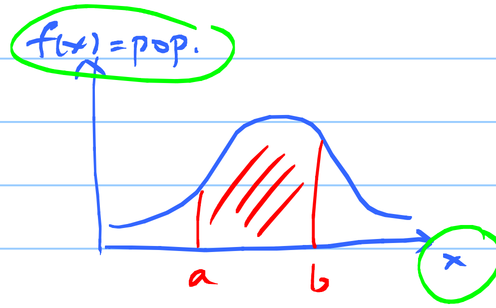
$$= pr(\;'' < z < ''\;) \Rightarrow \text{Table 1}$$



Compare with what we did in Ch 2:

$$pr(a<x<b) = pr\left(\frac{a-\mu_x}{\sigma_x} < \frac{x-\mu_x}{\sigma_x} < \frac{b-\mu_x}{\sigma_x}\right)$$

$$= pr(\;\cdots < z < \cdots\;)$$



$E[x] \quad E[\bar{x}]$

$\quad\quad\;\; \| \quad\quad\; \|$

Lots of means! $\bar{x}, \mu, \mu_x, \mu_{\bar{x}}$

sample mean

μ param of Normal

mean of pop./dist. x

mean of sampling distribution of $\bar{x}$

And, yes, in all of the above calculations of prob, we need to know the mu_x and sigma_x of population. So, this whole lecture does not seem to deliver on the promise of being able to determine mu_x and sigma_x of the population from a sample. For the delivery of that promise, wait for next lecture.

hw_lect15_1
Examine hw_lect7_1. In our new language, what you did there is to find the sampling distribution of the sample minimum (for samples of size n=3).
a) Revise the posted solution to find the sampling distribution of the sample mean for n=3.
b) Show that the mean (expected value) of that distribution is pi.

hw_lect15_2 (By R)
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) Repeat for the sample minimum.

Turn-in your 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_lect15_3 (By R)
a) write R code to take 5000 samples of size n=100 from an exponential distr. with parameter lambda=2, and make a qq-plot 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 sampling distr. of sample means is Normal, even when the pop. is not!

b) using the qq-plot, estimate the mean and std. dev. of the sampling dist. of sample means. Are they consistent with what you would expect from our formulas for the mean and standard deviation of the sampling distribution? show work.

$\mu_{\bar{x}} = \mu_x, \ \sigma_{\bar{x}} = \sigma_x/\sqrt{n}$

hw_lect15_4
A sample of size 36 from a Normal pop. yields the observed values xbar= 3.5 and s=1.
a) Under the assumption that mu_x = 2.5, and sigma_x = 2, what's the prob of a sample mean larger than the one observed?
b) Under the assumption that mu_x = 2.5, and sigma_x = 2, what's the prob of a sample mean smaller than the one observed?
c) Under the assumption that mu_x = 3.5, and sigma_x = 2, what's the prob of a sample mean larger than the one observed?
d) Under the assumption that mu_x = 3.5, and sigma_x = 2, what's the prob of a sample mean smaller than the one observed?
e) Now, suppose we know that sigma_x = 2, but we don't know mu_x. What is the observed 95% Confidence Interval for mu_x. Interpret it, in TWO ways.

postpone part e.

optional
hw_lect (By R)
Students are often suspicious of my claim that the E (i.e., distribution mean) of the ith element of a sample of size n is equal to the population mean, i.e. E[x_i] = E[x]. To convince yourself, write code to
- take 10^7 samples of size 50 from a normal distribution with mu = 2 and sigma = 3,
- select the 3rd element in each of the 10^7 samples, and store them in an array called x3.
- compute the mean of x3.
Convinced?!