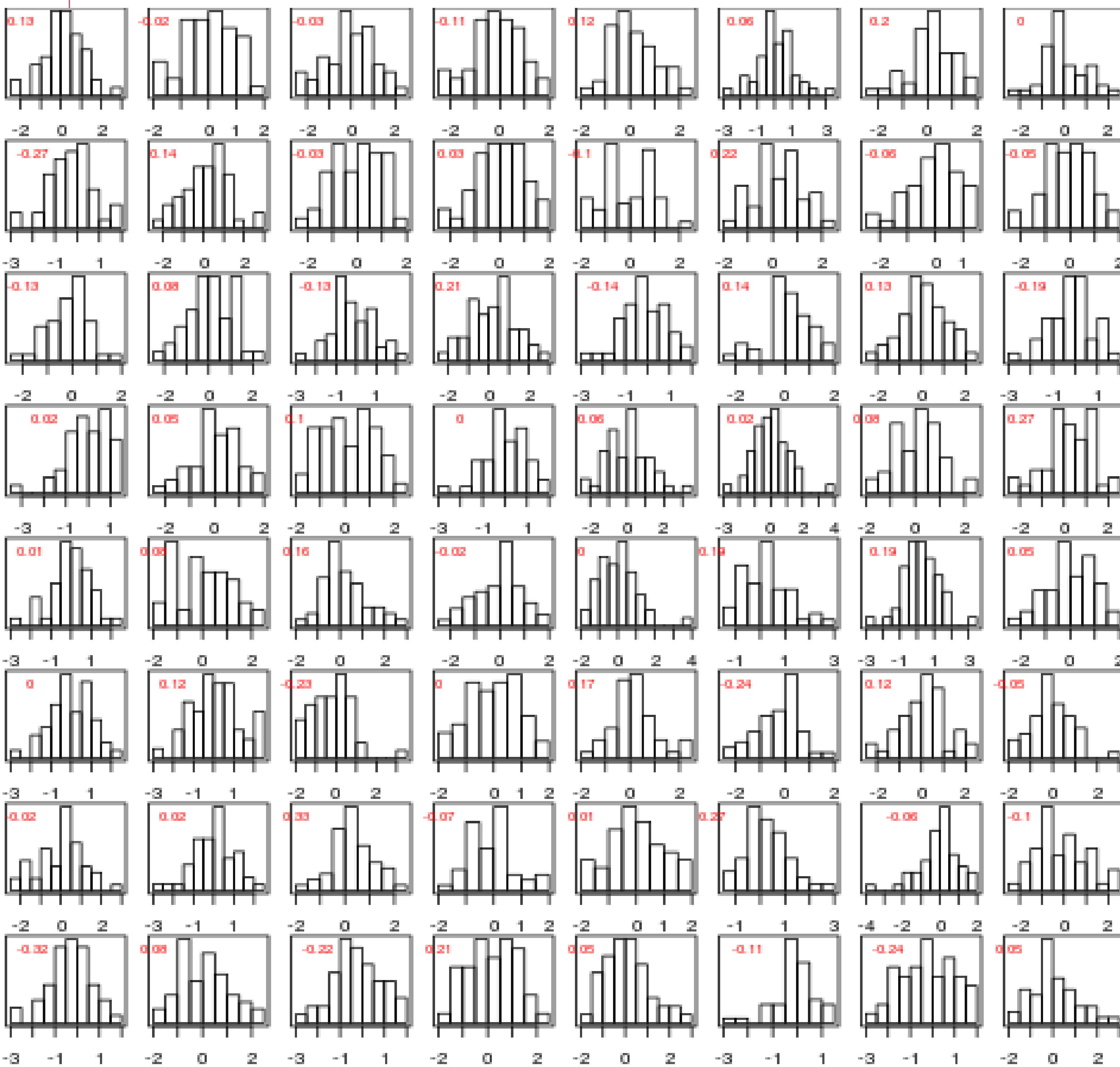
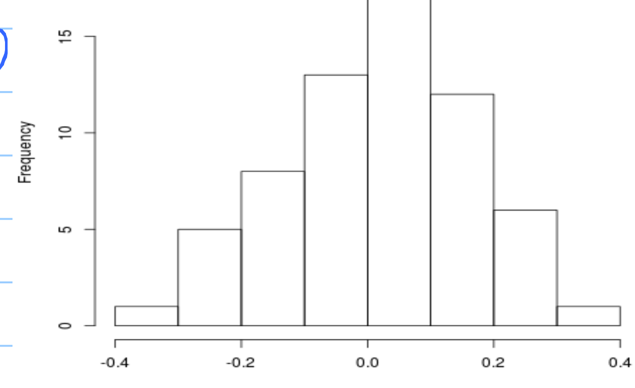


```
ntrial = 64
xbar = numeric(ntrial)
par(mfrow=c(8,8))
 for( trial in 1:ntrial ){
 x = rnorm(50, 0, 1)
 hist(x, breaks=10)
 xbar[trial] = mean(x)
 }
hist(xbar, main="")
```

← Try rexp(50,1)

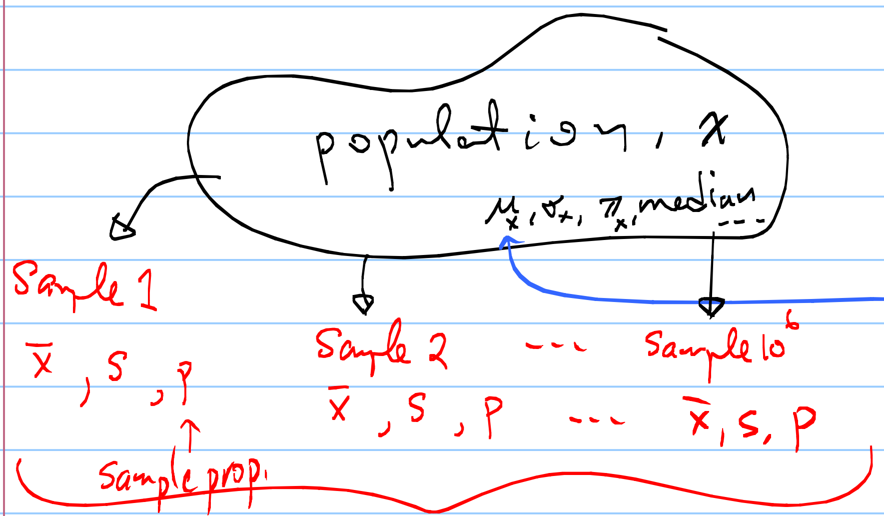


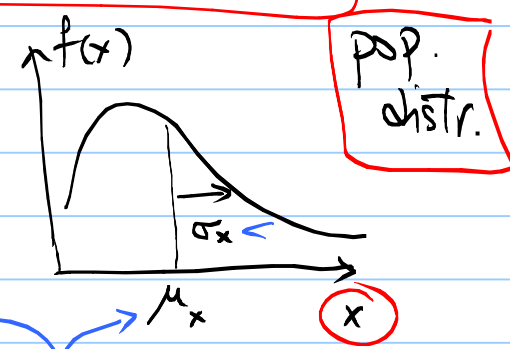

Q: What's $\bar{x}$ in each hist above?
What's the mean of the $\bar{x}$'s?

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

Lecture 18 (5.5, 5.6, ch.7)

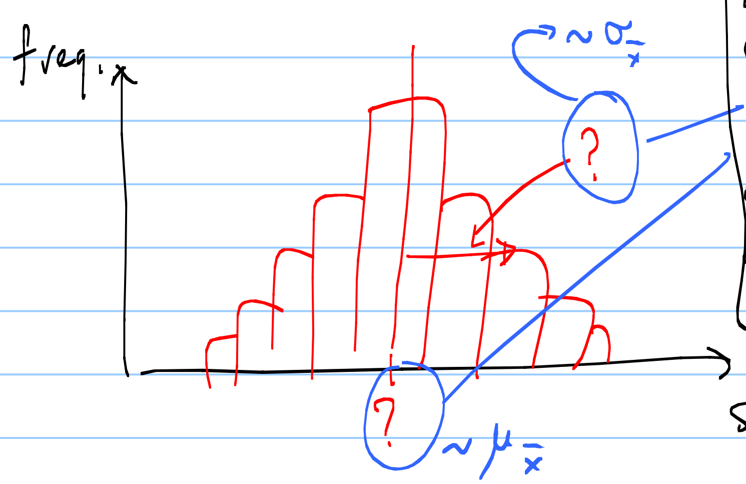
## Sampling Distribution : Extremely Important !!




- 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. $\mu$), 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 $\sigma^2_{\bar{x}}$ ):

If the population (ie. distribution) has mean $\mu_x$ and standard dev. $\sigma_x$, then

Mean of the Sampling distr. of sample mean :

Std. dev. " " " " " " " :

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

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

$\sigma_x$ ← pop. std. dev.

$n$ ← sample size

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 $\mu_x$ and $\sigma_x$

Of course, if you do know the pop. distr., then you can compute $\mu_x$, $\sigma_x$ 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

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

$$\Downarrow$$

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

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.

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

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

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

$\boxed{\text{Q1:}}$ Suppose we are taking samples of size 100 from a Normal dist. with params $\mu = 3$, $\sigma = 2$. In one such sample we observe a mean and std. dev. of 3.1 and 1.9, respectively. What is the std. dev. of the sample means?

a) 2   b) 1.9   $\boxed{\text{c) } \frac{2}{\sqrt{100}} = 0.2}$   d) $\frac{1.9}{\sqrt{100}} = 0.19$   e) none of the above

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 $\mu_x$ 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 (certain.) is our estimate of $\mu_x$.

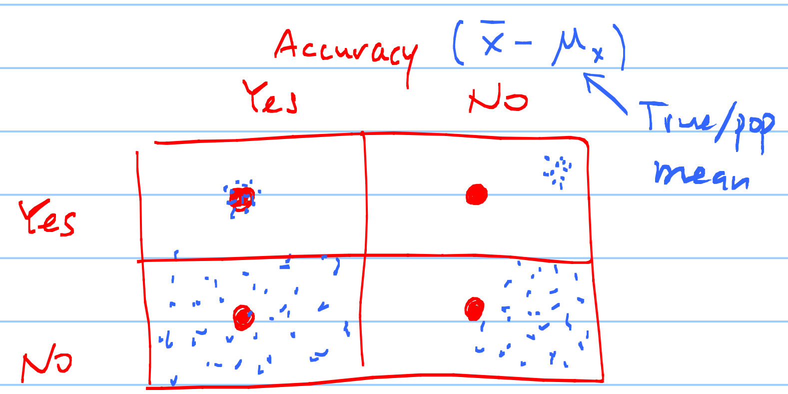
Note that $\mu_x, \sigma_x, \mu_{\bar{x}}, \sigma_{\bar{x}}$ are means and std. dev. of distributions, NOT of data. We are dealing with distributions, even though the thought exp. involved a hist.

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

---

FYI

$\bar{x}$ and $s_x$ are measures of Accuracy & Precision; and so $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$,




std dev. → precision

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

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

$\boxed{\text{Central Limit Theorem (CLT)}}$

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

---

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

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

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

1b) $\boxed{\text{If}}$ pop. distr. is not known, assume its $\mu_x, \sigma_x$ (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.

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

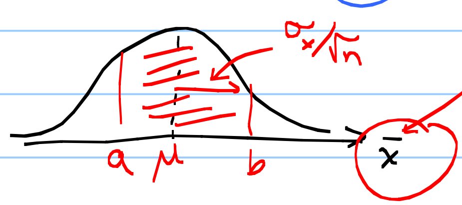



Table I

E.g. Suppose a sample of size 25 yields $\bar{x}_{obs} = 3$, $s_{obs} = 1$.
If the population is $N(\mu = 2, \sigma = 1)$, what's the prob.(*) of getting an even larger sample mean? $\to \mu_x = 2$, $\sigma_x = 1$

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

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

This small prob suggests that $\mu = 2$ is a bad assumption.
In fact, we may even guess that $\mu$ is greater than 2 (closer to 3)!
We will formalize these qualitative conclusions, below.

(*) Henceforth: "prob" = proportion of samples (of size n) taken from the population, in the long-run (e.g. out of $10^8$ samples)

---

hw-lect 18-1 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 <u>extreme events</u>, e.g. the biggest storms, the strongest earthquakes, the brightest stars, the smallest forms of life, etc.

hw-lect 18-2 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 - lect 18 - 3

A sample of size 36 from a Normal pop. yields $\bar{x} = 3, S = 1$.

a) Under the assumption that $\mu_x = 2.5, \sigma_x = 2$, what's The prob of a sample mean larger than the one observed.

b) Under the assumption that $\mu_x = 2.5, \sigma_x = 2$, what's The prob of a sample mean smaller than the one observed.

c) Under the assumption that $\mu_x = 3.5, \sigma_x = 2$, what's the prob of a sample mean larger than the one observed.

d) Under the assumption that $\mu_x = 3.5, \sigma_x = 2$, what's the prob of a sample mean smaller than the one observed.