Lecture 3 (Ch 1-2)

Sample vs. distr. ⇕ pop.

Histogram =  ⟷ $\boxed{\text{Sample data}}$

Distribution = 

Very different

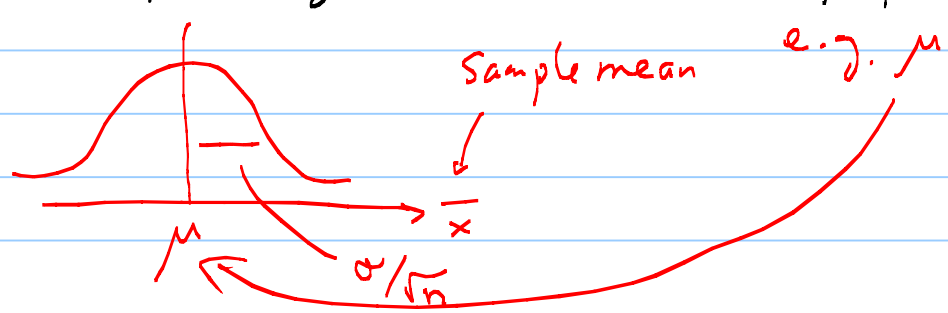
e.g. $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ = $\boxed{\text{mathematical function}}$

In stat. we use a distr. to represent a population.
A sample is a subset of the population.

An extremely important concept in stat.: $\boxed{\text{Sampling Distribution}}$
Intuitively: take a sample of size n, and compute a statistic, e.g. mean, median, std. dev. ... Then repeat a zillion times.
The hist. of the zillion statistic values is the empirical Samp. Dist.
The mathematical version of the above thought experiment gives the Samp. Dist.

One can have sampl. dist. of sample mean,
" median, or std. dev,

Each of these distributions tells us something about how accurately and precisely we can estimate pop. parameters.



Last time we wondered how to tell whether x (tip) effects y (Depth).

There are 2 ways:

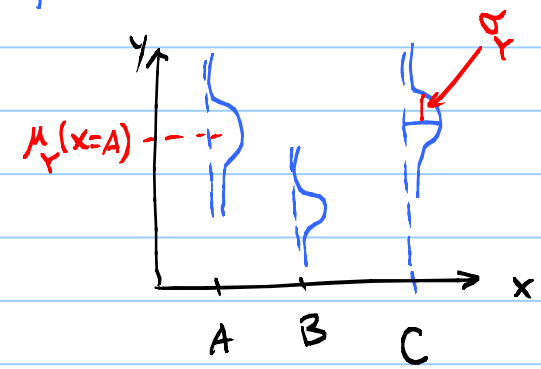
$\boxed{\text{One way:}}$ Assume the observed data (sample) come from some population that can be described by some distribution.

E.g. $$f_Y(y|x) = \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-\frac{1}{2}\left(\frac{y-\mu_Y(x)}{\sigma_Y}\right)^2}$$

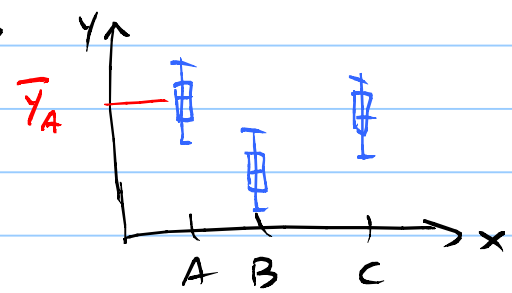
pop. params.

$x = \overbrace{A, B, C, \ldots}^{\text{many tips}}$

This is a prob. density function (pdf) with params $\mu_Y(x)$, σ_Y. We use such pdfs to model our data.



The data may be displayed similarly: →



Then we can do tests like

$$\begin{cases} H_0: & \mu_Y(x=A) = \mu_Y(x=B) \\ H_1: & \mu_Y(x=A) > \mu_Y(x=B) \end{cases}$$

Or we can build Conf. Int. (CI) for $\mu_Y(x=A)$, $\mu_Y(x=B)$, $\mu_Y(x=A) - \mu_Y(x=B)$,

← I'm assuming you know how to do These Things. But we'll do Them in Ch 2

As we saw, above, to compare box plots, variability is important. To do The tests, we need to know The variability of y, under H_0, ie. σ_Y. We'll learn to estimate it, later.

$\boxed{2^{nd} \text{ way}}$

Randomization test

Consider $\begin{cases} H_0: & x \text{ does not affect } y \\ H_1: & x \text{ affects } y \end{cases}$

This time, H_0, H_1 do not involve pop. params.

Suppose we randomly assign 2 tips to 8 coupons, and obs. the following data:

	Coupon	1	2	3	4	5	6	7	8
x	tip	A	B	A	A	B	B	A	B
y	Depth	13	11	12	9	7	8	6	5

Then $\bar{y}_A = \frac{1}{4}(13+12+9+8)$, $\bar{y}_B = \frac{1}{4}(11+7+8+5)$ $\quad \delta_{obs} \equiv \bar{y}_A - \bar{y}_B = \cdots$

To estimate the variability of δ under H_0, we use randomization.

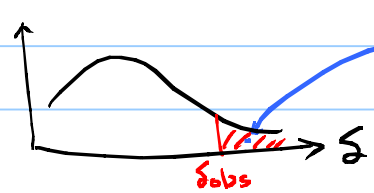
Under H_0 (ie. if H_0 = True), a different/random assignment of the tips (ie. the treatments) to the coupons (called Exp. Units) should not make any difference.

For example, under H_0, we are equally "likely" to "observe"

	Coupon	1	2	3	4	5	6	7	8
x	tip	B	A	A	A	B	B	A	B
y	Depth	13	11	12	9	7	8	6	5

$\Rightarrow \delta = \cdots$

The hist. of the $\binom{8}{4}$ δ values = randomization distr. of δ.



p-value. If small, then there is evidence against H_0, in support of H_1. Otherwise, data don't say anything!

The 2 tests give similar results, even though they are diff!

Most of what we do will use population-based test.

But even then, it helps to think about randomization.

Inference: Say something about pop. from sample. Dists

To do inference, we need notation:

Warning! You know all this, but here we focus on subscripts on $f(\ )$, $E[\ ]$, $V[\ ]$!

Let $Y =$ continuous (random) variable

Also $Y \neq y$, $X \neq x$ (for now)

→ The prob. density function (pdf) of Y is a function $f_Y(t)$ s.t.

$$f_Y(t) \geq 0 \quad , \quad \int_{-\infty}^{\infty} f_Y(t)\,dt = 1$$

Careful with notation!

$$\text{prob}(Y > a) = \int_{t=a}^{\infty} f_Y(t)\,dt = \int_{y=a}^{\infty} f_Y(y)\,dy = \int_{a}^{\infty} f(y)\,dy$$

→ The Expected Value of Y (under f_Y) is:

Carefully

$$\mu_Y \equiv E_Y[Y] = \int_{-\infty}^{\infty} t\, f_Y(t)\,dt \qquad = \int_{-\infty}^{\infty} y\, f_Y(y)\,dy = \int_{-\infty}^{\infty} y f(y)\,dy$$

→ The Exp. Value of Y^2 (ie. a function of Y) (under f_Y) is:

$$E_Y[Y^2] = \int_{-\infty}^{\infty} t^2\, f_Y(t)\,dt \qquad = \int_{-\infty}^{\infty} y^2 f_Y(y)\,dy = \int_{-\infty}^{\infty} y^2 f(y)\,dy$$

Carefully

→ $$\boxed{E_Y[g(Y)] = \int_{-\infty}^{\infty} g(t)\, f_Y(t)\,dt}$$ Defn. of $E_Y[g(Y)]$.

$V[\,]$

→ The Variance of Y (under f_Y) is:

$$\sigma_Y^2 \equiv \underset{\downarrow\, g(Y)}{V_Y[Y]} = \int_{-\infty}^{\infty} (\underset{\downarrow\, g(t)}{t} - \underset{\downarrow\, g(Y)}{E_Y[Y]})^2 f_Y(t)\, dt = E_Y[(Y - E_Y[Y])^2]$$

$$= E_Y[Y^2] - E_Y^2[Y]$$

Language: "Exp. Val. of Y" $\Longleftrightarrow$ "mean of f_Y" $\Longleftrightarrow$ "mean of Y"

What about 2 (or more) r.v.s? $\underbrace{f_{Y_1,Y_2}(t_1,t_2) = \text{----}}_{\text{joint pdf}}$

$$\iint_{-\infty}^{\infty} dt_1\, dt_2\, f_{Y_1,Y_2}(t_1,t_2) = 1 \quad \text{etc.}$$

Y_1 and Y_2 are $\boxed{\text{independent}}$ (defn.) iff $f_{Y_1,Y_2}(x_1,x_2) = f_{Y_1}(x_1)\cdot f_{Y_2}(x_2)$

→ The Covariance of Y_1, Y_2 is

$$\underset{Y_1,Y_2}{Cov}[Y_1, Y_2] = \int (t_1 - \mu_1)(t_2 - \overset{E_{Y_2}[Y_2]}{\mu_2})\, f_{Y_1,Y_2}(t_1,t_2)\, dt_1\, dt_2$$

$$= \underset{Y_1,Y_2}{E}[(Y_1 - \mu_1)(Y_2 - \mu_2)]$$

If Y_1 and Y_2 are indep. $\underset{\text{1-way.}}{\Longrightarrow}$ $Cov[Y_1,Y_2] = 0$ [ie. uncorrelated]

If $\boxed{Y = \text{discrete}}$, replace $\int$ with Σ.

Note: $\text{prob}(Y = a) = \begin{cases} \text{pmf of } Y \text{ at } Y = a, & \text{if } Y = \text{discrete} \\ 0, & \text{if } Y = \text{continuous} \end{cases}$

Again, pdf's/pmf's are mathematical quantities which are <u>a priori</u> <u>unrelated to data</u>.

They represent a population, not data/sample.

properties of $E[\,], V[\,]$

You will be showing some of these in your hw:

$$E_Y[c] = c, \qquad E_Y[cY] = cE_Y[Y],$$
$$V_Y[c] = 0, \qquad V_Y[cY] = c^2 V_Y[Y]$$

$$E_{Y_1,Y_2}[Y_1 \pm Y_2] = E_{Y_1}[Y_1] \pm E_{Y_2}[Y_2]$$

$$V_{Y_1,Y_2}[Y_1 \pm Y_2] = V_{Y_1}[Y_1] + V_{Y_2}[Y_2] \pm 2\,\mathrm{Cov}_{Y_1,Y_2}[Y_1,Y_2]$$

If Y_1 and Y_2 are indep, then $\mathrm{Cov}[Y_1,Y_2]=0$ (not vice versa)

and $E_{Y_1,Y_2}[Y_1 Y_2] = E_{Y_1}[Y_1] \cdot E_{Y_2}[Y_2]$

Example

If $f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}$

Change of var. $z = \frac{t-\mu}{\sigma}$

Then $\mu_Y \equiv E_Y[Y] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}}\, t\, e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt = \cdots = \mu \cdot 1$

$$\sigma_Y^2 \equiv V_Y[Y] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{\ }} (t-\mu_Y)^2\, e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt = \cdots = \sigma^2 \cdot 1$$

I.e. $Y \sim N(\mu, \sigma^2) \implies \mu_Y = \mu, \quad \sigma_Y = \sigma$

This is why I put subscripts on μ_Y, σ_Y; to distinguish them from μ, σ params of Normal.

We can find μ_Y, σ_Y for any pdf/pmf just Normal.

E_Y and V_Y tell us about the location and "width" of f_Y.

typical value of Y — typical deviation in Y

From now on, I'll be sloppy with Y, y, ... They are all r.v.'s.

On the sample/data side, the analogous quantities are

Sample mean $= \bar{y} = \frac{1}{n}\boxed{\sum_i y_i} \equiv y_\bullet$ = sum of y's, ie. Total y.

Sample variance $= s^2 = \frac{1}{n-1}\boxed{\sum_{i=1}^{n}(y_i-\bar{y})^2} \equiv SS$ = Corrected sum of squares.

$\bar{y}$ and s^2 are random variables, so let's find their E, V:

$\rightarrow E[\bar{Y}] = E\left[\frac{1}{n}\sum_i^n Y_i\right] = \frac{1}{n}\sum_i \underbrace{E[Y_i]}_{E[Y] \text{ for all } i;\ Y_i = \text{identically distributed}} = \mu_Y \Rightarrow \boxed{E[\bar{Y}] = \mu_Y}$

($\bar{Y}$ = unbiased estimator of μ_Y)

"μ_Y = pop. mean

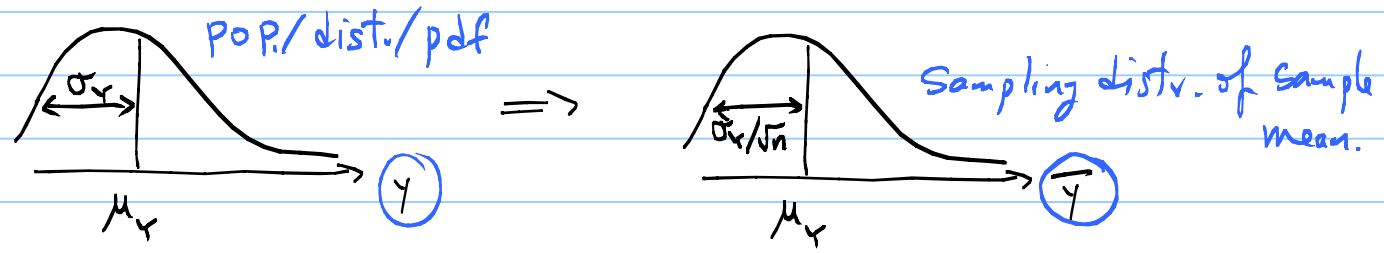
$\rightarrow V[\bar{Y}] = V\left[\frac{1}{n}\sum_{i=1}^n Y_i\right] = \frac{1}{n^2}\sum_i \underbrace{V[Y_i]}_{V[Y] \text{ for all } i} + 0 = \frac{\sigma_Y^2}{n} \Rightarrow \boxed{V[\bar{Y}] = \frac{\sigma_Y^2}{n}}$

"σ_Y^2 = pop. var.

$Cov[Y_i, Y_j] = 0 \quad i \neq j$

Y_i = independently dist.

Pictorially, for iid Y_i:



hw-lect3-1

Consider the problem of testing the following pair of hypotheses:
H0: x is related to y
H1: x is not related to y

The parametric approach to testing those hypotheses might look like:
H0: mu_x = mu_y
H1: mu_x > mu_y
where mu_x and mu_y are the mu parameters of two normal distributions for x and y, respectively. Suppose the sigma parameter of the two distributions are known to be sigma_x and sigma_y, respectively. This is the famous two-sample, 1-sided test that we are all familiar with from basic stats. We know that an important quantity to look at is the area under the normal distribution with parameters mu=0, sigma = sqrt(sigma_x^2/n + sigma_y^2/n), where n is the sample size.

Suppose this is our observed data:

```
set.seed(123)
nsample = 100
x_obs = rnorm(nsample,0,1)
y_obs = rnorm(nsample,0,1)
delta_obs = mean(x_obs) - mean(y_obs)
```

a) Using pnorm() find the p-value. Report the number you get.

Now, the randomization test: If the populations from which x and y have been drawn are truly the same, then in theory, we should be able to to switch each case in x with a case in y, and still get the same mean for the two samples. This way we can build the sampling distribution of delta, and then the area under that sampling distribution, to the right of the observed delta will be something like a p-value.

There are many ways of switching things, and we can get the delta for each of them. The hist of the resulting delta values will be the sampling distribution of delta. However, instead of switching things, it's easier to join (pool) the two observed samples together into one array, take a random sample of size 4 from that array and assign it to the x-group, and assign the unsampled elements of that array to the y-group. Etc. (By the way, when *all* possible permutations are examined, one says that one has done an "exact test." If one considers only a sample of all possible permutations, then one says that one has performed a "monte carlo test.") For example, if we have an array of 8 integers, z, we can sample 4 of them this way:

```
z = c(1:8)
index = sample(z, 4, rep = F)
```

Then, for example, the sampled z values can be selected this way:

```
z[index]
```

In R, it's easy to select the "unselected" part this way:

```
z[-index]
```

b) So, knowing this trick, combine x_obs and y_obs into a single array, using c(), called pool. and then take ntrial=1000 samples of size 100 from pool. Use the above trick to separate the x and y samples. Then for each of the ntrial runs, compute delta = mean(x) - mean(y).
IMPORTANT: First type in set.seed(123)

c) Plot the histogram of delta, and find the p-value.

hw-lab 3-2

Suppose $X \sim N(\mu, \sigma)$, ie. $f_X(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2}$ $\quad -\infty < t < \infty$

Use only the defn. of $E_Y[g(Y)]$ given above, to find

[use $\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}x^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}x^2} dx = 1$, $\int_{-\infty}^{\infty} x e^{-\frac{1}{2}x^2} dx = 0$]

This is important! For example do NOT use properties of $E[\,]$, even if you derive them from the defn.

a) $\mu_X \equiv E_X[X]$

b) $E_X[X^2]$

c) Now, assume $X \sim N(0,1)$. Find $E_{X^2}[X]$ using the chisqd pdf.

Hint: if $X \sim N(0,1)$, then $X^2 \sim$ chisqd with df=1 where the pdf of chisqd with df=k is given by

$$f_{X^2}(t) = 2^{-\frac{k}{2}} \frac{1}{\Gamma(\frac{k}{2})} t^{\frac{k}{2}-1} e^{-\frac{1}{2}t} \qquad t > 0$$

Hint: $\Gamma(x) = (x-1)! = (x-1)(x-2)! = (x-1)\Gamma(x-1)$

Hint: $\int_0^{\infty} f_{X^2}(t)\,dt = 1$ for all k.

d) Find $E_{X^2}[X^2]$ using the chisqd pdf.

e) It can be shown that if $Y = X^2$, then $f_Y(t) = \frac{1}{2\sqrt{t}}\left(f_X(\sqrt{t}) + f_X(-\sqrt{t})\right)$ $\quad t \geq 0$.

Find $E_{X^2}[X^2]$ (ie. $E_Y[Y]$), but this time using this pdf of Y, and show that it's equal to $E_X[X^2]$.

Hint: use change of variables, and be careful with integration limits.

hw- lect 3-3

Starting from the defn. of $E_{Y_1,Y_2}[\,]$, $V_{Y_1,Y_2}[\,]$, and $\mathrm{Cov}_{Y_1,Y_2}[\,,\,]$ show

$$V_{Y_1,Y_2}[Y_1 \pm Y_2] = V_{Y_1}[Y_1] + V_{Y_2}[Y_2] \pm 2\,\mathrm{Cov}_{Y_1,Y_2}[Y_1, Y_2]$$

Don't skip any steps!