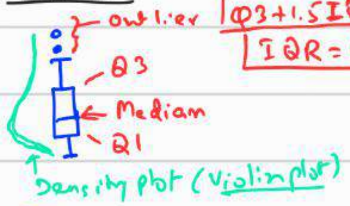
- Bar Chart - Categorical
- histogram - Numerical / Continuous
- Density Plot - Smoothed histogram (kernel)

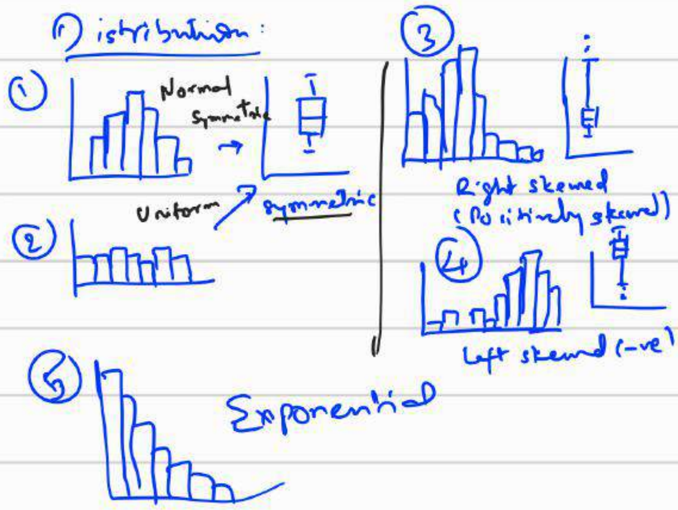
Box Plots:

$\boxed{Q3 + 1.5\,IQR}$

$\boxed{IQR = Q3 - Q1}$

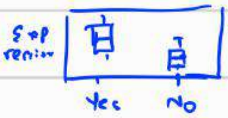


Distribution:



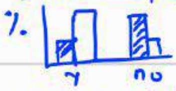
[2 Variables]:

1. 1 cat, 1 numeric side by side boxplot



2. 2 cat

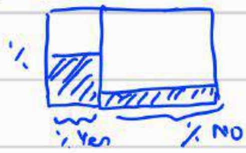
a) side by side bar chart



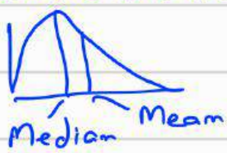
b) Stacked

c) Mosaic

d) scatter:



- Trimmed mean: Mean after removing top/bottom α% of data.
- Median - Not sensitive to outliers.
  - Non parametric measure



- Standard Dev.: $\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$
- Average Abs. Dev. $\sqrt{\dfrac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n-1}}$

why $(n-1)$

If we have n obs. we loose one d.f for estimation.

For eg. $n=1$ & $x_1 = 86$, S.D = ? no S.D as we have only one obs.

let's say $n=4$ and given a statistic $\bar{x}=80$, we can choose X_1, X_2, X_3 freely but X_4 has to be some number to make $\bar{x}=80$. so d.f = 3

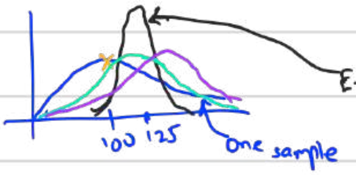
Normal Dist^n : 68 / 95 / 99.7

$$\boxed{Z = \frac{x-\mu}{\sigma}}$$

Inferential statistics -

Inference of population from sample.

Sampling distribution - Theoretical set of ALL POSSIBLE statistics (eg. $\bar{x}$) we could get. Eg. taking sample size of 25 over and over again. Assume pop. mean $\mu = 125$, $\sigma = 20$, one sample mean = 100, $n = 25$



Expected value of $\bar{x} = \mu = 125$

$\underline{S.D \text{ of } \bar{x}}$ = Standard Error $(SE_{\bar{x}}) = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{25}} = 4$

on average how far estimated $\bar{x}$ deviate from μ and as it is $\frac{\sigma}{\sqrt{n}}$ ie, as sample size increases the S.E is smaller.

S.D of $\bar{x}$: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

S.E of $\bar{x}$: $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

s is sample S.D

Some Properties:

① $Var(ax) = a^2 Var(x)$ a is constant

② $Var(x_1 + x_2) = Var(x_1) + Var(x_2)$ if x_1, x_2 independent

x_i has mean μ, var $= \sigma^2$

$\bar{x} = \frac{x_1 + x_2 \dots + x_n}{n}$

Now,

$$Var(\bar{x}) = Var\left(\frac{x_1 + x_2 + \dots x_n}{n}\right)$$

$$= \frac{1}{n^2} Var(x_1 + x_2 + \dots x_n)$$

$$= \frac{1}{n^2}\left[Var(x_1) + Var(x_2) + \dots Var(x_n)\right]$$

$$= \frac{1}{n^2}\left[\sigma^2 + \sigma^2 + \dots + \sigma^2\right]$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \Rightarrow S.D_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Confidence Interval

Let's look at a skewed dist^n $\mu = 125$, $\sigma = 20$; Sample $\bar{x} \approx$ Normal $\mu_{\bar{x}} = 125$, $\sigma_{\bar{x}} = \frac{20}{\sqrt{25}} = 4$ $n = 25$

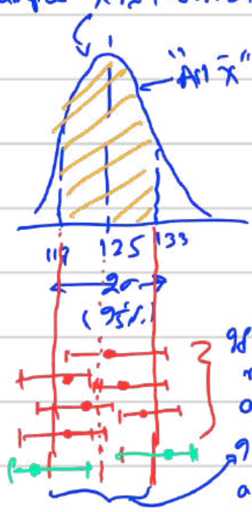
→ 95% of time the $\bar{x}$ will be within $\mu \pm 2\, SD_{\bar{x}}$ or $\mu \pm 2\sigma/\sqrt{n}$

→ conversely 95% time the μ will be within $\bar{x} \pm 2 \cdot SD_{\bar{x}}$ or $\boxed{\bar{x} \pm 2\frac{\sigma}{\sqrt{n}}}$

In reality we don't know σ, replace with s

$\boxed{\bar{x} \pm t \cdot \frac{s}{\sqrt{n}}}$ where $\frac{s}{\sqrt{n}}$ is $S.E_{\bar{x}}$

"t" instead of 2

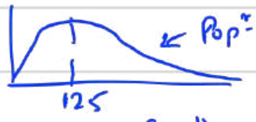


If we draw a sample n = 25, it could be any of these points

95% times in this region and true mean lies within

Hypothesis Testing:

What's the prob. of $\bar{x} \geq 135$?

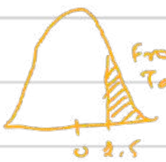


let's standardize

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{135 - 125}{4} = 2.5$$

This mean

$\bar{x} = 135$ is 2.5 S.D above the mean



→ Pop^n Mean 125, Prob. of estimating $\bar{x} \geq 135$ is 0.6%.

Eg. For smokers we believe that $\mu > 125$; if 125 is for healthy pop^n.

We take a sample ($n = 25$) of smokers – Their mean $\bar{x} = 135$ & $s = 20$.

H_0 : $\mu_{smokers} = 125$

H_1 : $\mu_{smokers} > 125$

For this we use t- dist^n (same as Z dist^n)

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

(can be thought of as Normal dist^n for samples)

Now in this scenario 0.6% is p-value

⇒ Assuming "Pop^n mean is 125"

t-dist^n:

used instead of z-dist^n b/c don't know pop^n std. dev. and must use sample std. dev. to estimate std. Error.

Sample std. dev. is not a true estimate of $\sqrt{\sigma}$

So as $n \to \infty$

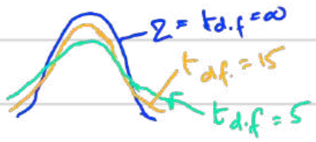
$\sqrt{s} \to \sqrt{\sigma}$

Conf. interval

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

Hyp. testing

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



Confidence Interval for Mean

Eg. Mean BMI of pop^n $\mu = 27.2$

take sample $n = 16$, $\bar{x} = 25.2$, $s = 5$

$$\bar{x} \pm t_{n-1}^* \frac{s}{\sqrt{n}}$$

95% C.I
2-sided

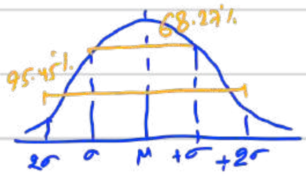
$$= 25.2 \pm 2 * \frac{5}{\sqrt{16}} = 25.2 \pm 2.5 \quad (22.7, 27.7)$$

$t_{15} = 2$ (just for example)

If we take repeated samples of $n = 16$
for every 100 C.I, about 95% will have the true mean within the intervals.

one-sided 95%,
we interested in upper bound

$$\bar{X} + (1.64)\frac{5}{\sqrt{16}} = 25.2 + 1.64\left(\frac{5}{\sqrt{16}}\right) = 27.25$$



this qty refers to value
That says 95% of observation lie below 1.64 s.d above the mean or less

Controlling the margin of error:

$$\bar{X} \pm \underbrace{t\left(\frac{s}{\sqrt{n}}\right)}_{\text{margin of error}}$$

Reducing ME : $\downarrow t$ - reducing t - lower CI - ✗
$\downarrow s$ - can't be reduced
$\uparrow n$

Suppose we want M.E 0.5 i.e. $\bar{x} \pm 0.5$

$$n = \left(\frac{t * s}{ME}\right)^2 = \frac{2 \times 5}{0.5} = 400$$ | value from previous page

If you are planing expt. - to get s - Literature / - Small pilot study

Bootstrapping & Resampling:

Parametric / large "n"

Pop^n — sample #1 $\bar{X}_1$ — sample #2 $\bar{X}_2$ — sample #3 $\bar{X}_3$

sampling dist^n $\cong$ Normal
s.D of $\bar{X} = \sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$

- Bootstrap : Instead of large sample.
  & when SE estimate is difficult.

sample → $\bar{X}$
Re sample with replacement $\bar{X}_2^*$
" " " $\bar{X}_3^*$
$\vdots$
$\bar{X}_B^*$
} Bootstrap Sampling dist^n

Eg. 60, 75, 80, 85, 90, with $\bar{x} = 78$, $s = 11.51$, $S.E = \frac{11.51}{\sqrt{5}} = 5.15$

Random sample #1 : Pick a sample 75, then another w. replacement 90, ... we get
75, 90, 80, 90, 85 $\quad \bar{x}_1^* = 84$
#2 $\quad$ 85, 60, 75, 85, 60 $\quad \bar{x}_2^* = 73$
$\vdots$

$SE_{\bar{x}}^* = 5.57$ (quite close to 5.15)

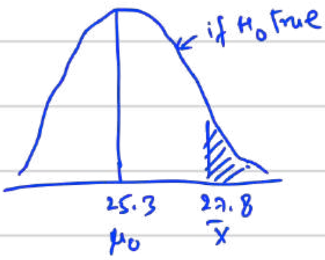
One-sample t-test:

known Mean BMI USA 2008 = 25.3 .. has it increased

2018 $n=25$, $\bar{x}=27.8$, $s=6$

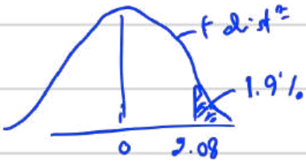
H_0 $\mu_{2018} = 25.3$ – why H_0 is stated this way? because we know it is true. if we make H_A as null we don't what is true, what do we expect

H_A $\mu_{2018} > 25.3$



$$t_{statistic} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{27.8-25.3}{6/\sqrt{25}} = 2.08$$

27.8 is 2.08 stand. errors above what we'd expect if H_0 true.



pvalue 0.019 ie, if H_0 true, prob. of $\bar{x} \geq 27.8$ is $\leq 1.9\%$

so H_0 not true

95% CI $\quad 27.8 \pm 2\left(\frac{6}{\sqrt{25}}\right) \to (25.4, 30.2)$

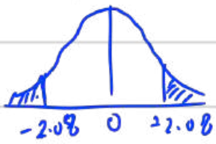
One Vs. Two sided: t. test: $\quad t_{STA} = 2.08$

One sided: H_A $\mu_{2018} > 25.3$

here we're looking at prob. of getting estimate 2.08 or more above 25.3 (μ_0)

Two sided: H_A: $\mu_{2018} \neq 25.3$

Prob. of getting estimate that is 2.08 or more std. Error away from μ_0



This will be 3.8%.

What if One-sided p-value is 3%, then two would be 6%, – so reject H_0

well depends on the problem.

Hypothesis test & CI:

H_0: $\mu = \mu_0$ $\quad H_A$: $\mu \neq \mu_0$

$$t_{STA} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} \qquad CI: \bar{x} \pm t * \frac{s}{\sqrt{n}}$$

↳ gives pvalue – α