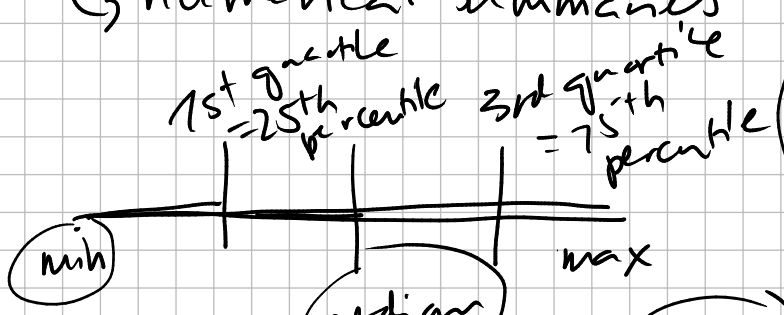
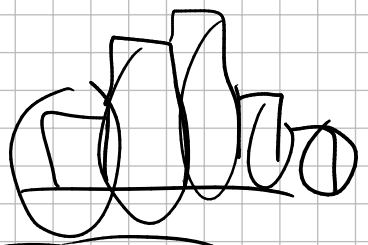
histogram, box plot

↳ numerical summaries

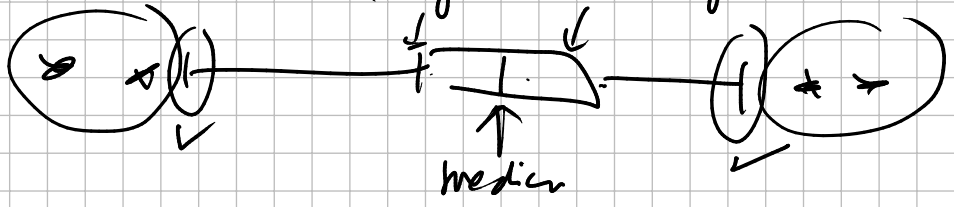


- lower limit
- upper limit
- count



data plenty of observations

reduce dimensions



1, 2, 3, 4, 5

1, 2, 3, 4

1000

1, 2, 3, 4, 5

2.5

continuous variables

histograms / box plot / some of numerical summaries

might not be useful for categorical variables } discrete variables
dummy variables

mean (average)

$$\frac{1}{n}\sum_{t=1}^{n} X_t$$

X_t variable indexed by t

t → subscript

n ← ending point

$\sum$ = summation

$t=1$ → starting point

PERNUM

PERWT

$$\sum_{t=1}^{n} X_t = X_1 + X_2 + X_3 + \cdots + X_n$$

ending point

$t=1$ $t=2$ $t=3$ $\cdots$ $t=n$

might represent time

t	X_t
1	X_1
2	X_2
3	X_3
$\vdots$	$\vdots$
n	X_n
Σ	$\sum_{t=1}^{n} X_t$

$$\text{mean} = \bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t \quad = \text{simple average}$$

$$\sum_{t=3}^{n-2} X_t = X_3 + X_4 + \cdots + X_{n-2}$$

moving average

trimmed mean

Survey:

$$\sum_{t=3}^{6} (-1)^t X_t = (-1)^3 X_3 + (-1)^4 X_4 + (-1)^5 X_5 + (-1)^6 X_6$$

$$= (-1)^3 \cdot 3 + (-1)^4 \cdot 4 + (-1)^5 \cdot 5 + (-1)^6 (6) = 2$$

$X_t = t$

$$\sum_{t=3}^{6} X_t = X_3 + X_4 + X_5 + X_6 = 2 + 2 + 2 + 2 = 8$$

Suppose $X_t = 2$, where c is some constant, for all t

$$\sum_{t=3}^{6} X_t = \sum_{t=3}^{6} 2 = 2\sum_{t=3}^{6} 1 = 2(1+1+1+1) = 8$$

$$\sum_{t=3}^{6} X_t = 4 \cdot c \quad \text{if } X_t = c \text{ for all } c$$

$$\sum_{t=3}^{6} X_t = X_3 + X_4 + X_5 + X_6 = 2y_3 + 2y_4 + 2y_5 + 2y_6$$

$$= 2(y_3 + y_4 + y_5 + y_6)$$

$$= 2\sum_{t=3}^{6} y_t$$

Suppose $X_t = 2y_t$

In other words, $\sum_{t=3}^{6} 2y_t = 2\sum_{t=3}^{6} y_t$

$$\sum_{t=3}^{6} (X_t)^2 = X_3^2 + X_4^2 + X_5^2 + X_6^2$$

NOTE: $\left(\sum_{t=3}^{6} X_t\right)^2 = \left[X_3 + X_4 + X_5 + X_6\right]^2$

$\leq$ generally, not equal!

where in the histogram can you find the mean?

$$\bar{X} = \frac{1}{n}\sum_{t=1}^{n} X_t$$

$$0 = \left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) - \bar{X}$$

any X_t !

$$= \left(\frac{1}{n}\sum_{t=1}^{n} X_t\right) - \left(\frac{1}{n}\sum_{t=1}^{n} \bar{X}\right)$$

$$\bar{X} \quad \bar{X} \quad \cdots$$

$t=1 \quad t=2$

$n \cdot \bar{X}$

$$0 = \frac{1}{n}\sum_{t=1}^{n} X_t - \frac{1}{n}\sum_{t=1}^{n} \bar{X} = \frac{1}{n}\sum_{t=1}^{n} (X_t - \bar{X})$$

average of $(X_t - \bar{X})$'s $= 0$

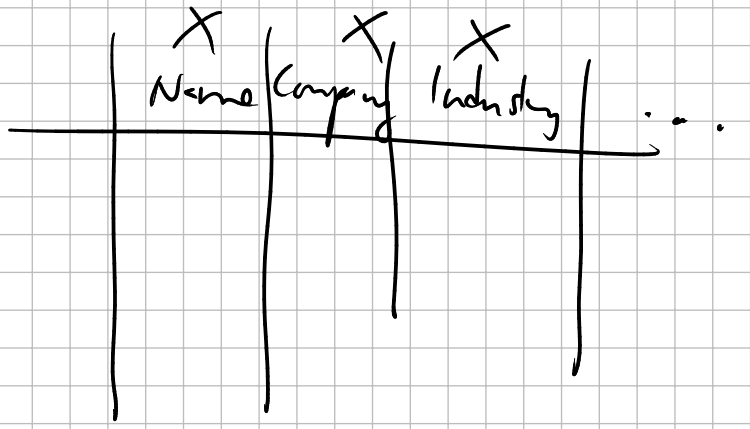
$(X_t - \bar{X})$ — deviation from the mean

how every observation of X differs from $\bar{X}$

negative deviations $X_t < \bar{X}$

$X_t > \bar{X}$ positive deviations

sd( ) std dev of numerical variable



sd(execComp$TotalCompMil) =
sd(execComp$SalaryThou) =
apply

std dev $= \sqrt{\frac{1}{n}\sum_{t=1}^{n}(X_t-\bar{X})^2}$ or sometimes $\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}(X_t-\bar{X})^2}$

$(X_t-\bar{X})^2$ $(X_t-\bar{X})^2 \neq (0.01)^2$

It does not make sense to propose a measure of spread based on $\frac{1}{n}\sum_{t=1}^{n}(X_t-\bar{X})$!

$\underbrace{\quad}_{\text{always } 0}$

absolute deviation

$|-3|=3$
$|3|=3$

$\frac{1}{n}\sum\widetilde{|X_t-\bar{X}|}$

alternative measure of spread

$$\text{variance} = \frac{1}{n}\sum\left(X_t - \bar{X}\right)^2$$

changes the measurement of X_t (units)

X_t measured in dollars

$\left(X_t - \bar{X}\right)^2$ measured in dollars^2

mean(execComp$TotalCompMil) → see what happens!

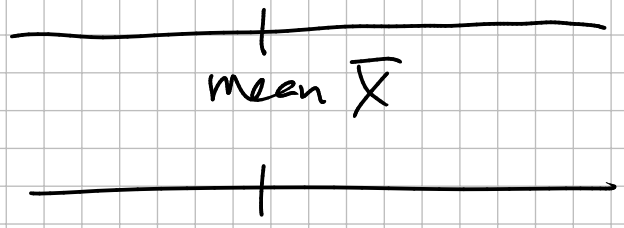
$$-1\ \begin{array}{r} 461300 \\ -696798 \\ \hline -255498 \end{array}\ \text{dollars} \qquad \begin{array}{r} 461.3 \\ -696.7 \\ \hline -255.4 \end{array}\ \text{thousand dollars}$$

$$\frac{X_t - \bar{X}}{sd} = \text{number}$$ no unit/unitless

$$X_t - \bar{X} = \text{number} * sd \Rightarrow X_t = \bar{X} + \text{number} * sd$$

± 3

$$\frac{X_t - \bar{X}}{sd}$$

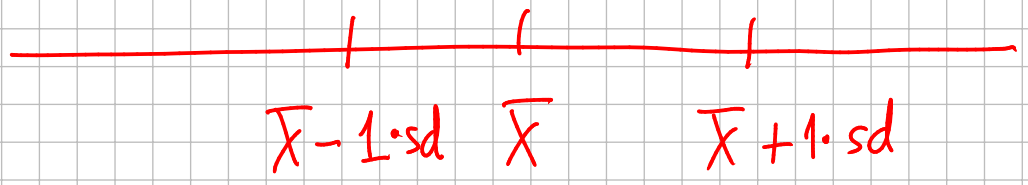


mean of $\left(\frac{X_t - \bar{X}}{sd} \right)$

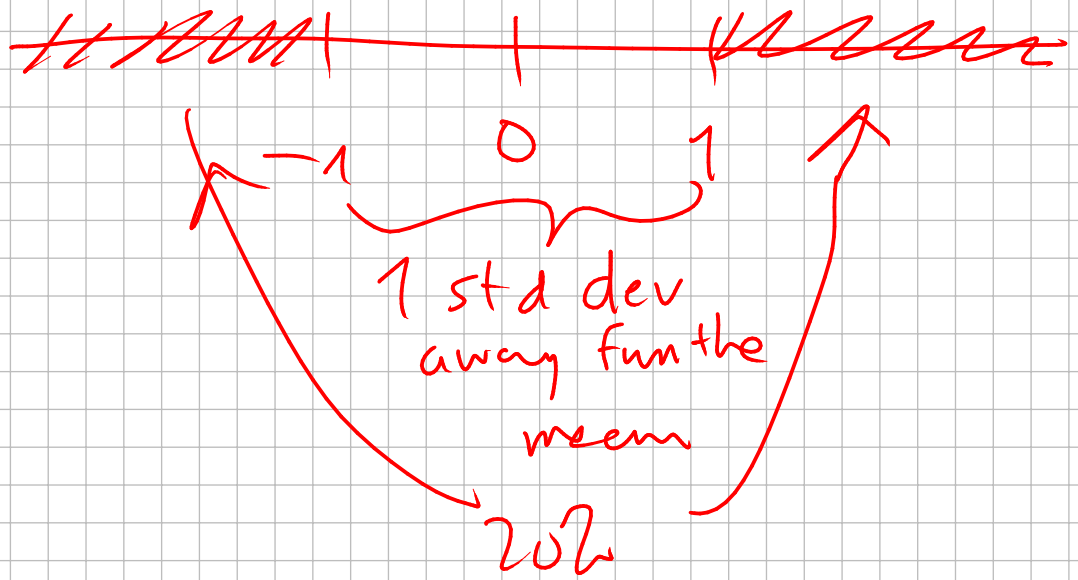
mean of $\left(\text{abs}(z_{Sal}) > 1 \right)$? → proportion

150 z-scores

z-scores	z_{Sal}	$\text{abs}(z_{Sal}) > 1$
1.2	1	TRUE
0.5	0	FALSE
0.6	0	FALSE
−1.7	1	TRUE
1.8	1	TRUE
...	...	...



original scale
dollars



Standardized scale
no units

6.29e+02 in R

$6.29 * 10^2$ scientific notation
$= 629$

$X_t \rightarrow X_t - \bar{X}$ translation

$X_t \rightarrow \frac{X_t - \bar{X}}{sd}$ combination of translation & scaling

$F = \frac{9}{5}C + 32$

$X_t \longrightarrow \log X_t$

natural logarithmic transformation

weird for economics algebra

$\log X_t = \log_e X_t$

$\log X_t = \log_{10} X_t$

↳ strange you note!

$X_t \longrightarrow X_t^2$

$X_t \longrightarrow (X_t - \bar{X})^2$ } nonlinear transformations

↳ $(X_t - \bar{X})(X_t - \bar{X})$

$\log_b x = ?$

↳ exponent

$\log_b x = y \iff b^y = x$

$\log_b x_2 + \log_b x_1 = \log_b (x_1 \cdot x_2)$

$$\log_b x_2 - \log_b x_1 = \log_b \left(\frac{x_2}{x_1}\right) \quad \Big| \quad \log_b x^a = a \log_b x$$

$\log_{(10)} 50 = y$ if $\log_{10} 5 = 0.7$

$\log_{(10)} 50 = 10^2$ $\log_{10} 5 = 0.7 \Rightarrow 10^{0.7} = 5$

$10^y = 50 \Rightarrow (10^2)^y = 50 \Rightarrow 10^{2y} = 50$

$\Rightarrow \log_{10} 10^{2y} = \log_{10} 50$

$\log 0$ does not exist

$\ln R$ $-\infty$ INF

$10^? = 10$

$2y \log_{10} 10 = \log_{10} 50$

$\log_{10} 10 = 1$

$2y = \log_{10} 50$

$y = \frac{\log_{10} 50}{2} = 1.7/2 = 0.85$

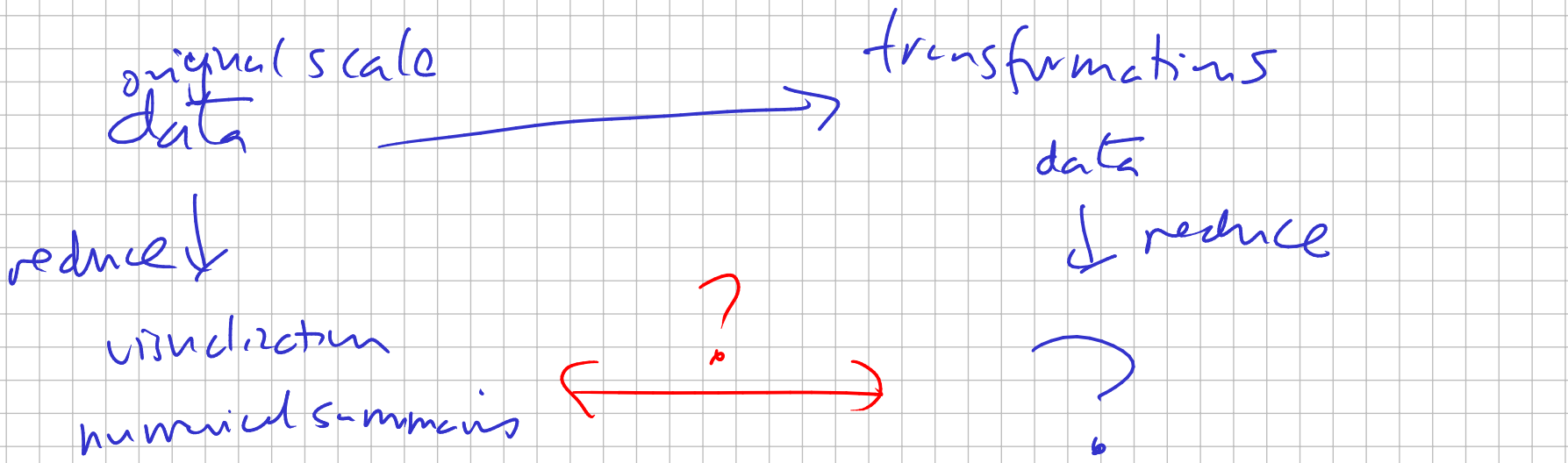
$\log_{10} 50 = \log_{10}(5 \cdot 10)$

$= \log_{10} 5 + \log_{10} 10$

$\log_{10} 5 = 0.7$ $\log_{10} 10 = 1$

NOTE

$\log_b (x_1 + x_2) = ? \neq \log_b x_1 + \log_b x_2$



a^x a^x

10^6 = 10^6

+e06 $*10^6$ (scientific notation)