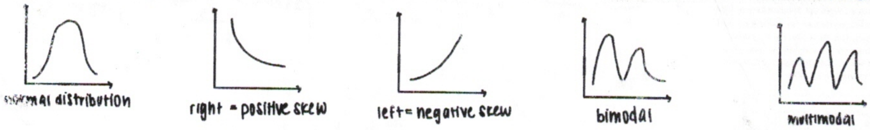
Qualitative - categorical (not measured) / quantitative variables (numerical) - discrete = finite , continuous = not an exact value



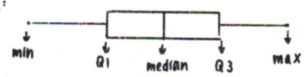
mean > median > mode = skewed right
mean < median < mode = skewed left
mean = median → normal distribution

★ Skew determined by tail

★ leading zeros do not count to sig figs. trailing does

sample standard deviation: $s = \sqrt{\dfrac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$ $\bar{x}$ = mean, N = number of observations ★

box plots:

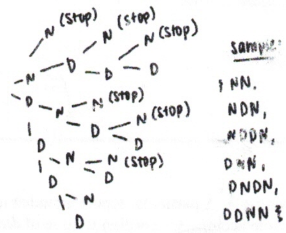


IQR = Q3 - Q1

lower fence = Q1 - 1.5 x IQR
upper fence = Q1 + 1.5 x IQR
] outliers occur outside the lower and/or upper fence ★

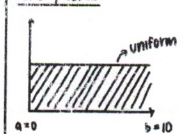
P(A) = # A / total # of possible outcomes
P(not A) = 1 - A
A^c = outcomes in the sample space but not in event A
A∪B = in either A or B
↳ P(A∪B) = P(A) + P(B) = mutually exclusive
↳ P(A∪B) = P(A) + P(B) - P(A∩B) = not mutually exclusive
A∩B = in both A & B
↳ P(A∩B) = 0 → mutually exclusive
↳ P(A∩B) = common / total → not mutually exclusive
conditional "given" - P(A|B) = (# A∩B) / total B
independent events = P(A∩B) = P(A) · P(B)
dependent events = P(A∩B) = P(A) · P(B|A)

tree diagram - terminates at a nondefective batteries (max 4 runs)



μ_x (average) = Σ (x)(p(x))
σ_x^2 (variance) = Σ $(x - \mu_x)^2 p(x)$ → √ to find standard deviation (σ_x)

density curve:



height = $\frac{1}{b-a}$ P(#₁ < x < #₂) = P(a < x < b) = P(x < b) - P(x - a)

converting from p → z = $\dfrac{x - \mu}{\sigma}$ (x: given value, μ: mean, σ: standard deviation)

the sampling distribution of a sample mean

converting p → z : $\dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}}$ (μ: mean of the Q; $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}}$, σ: standard deviation, n: sample size)

Exercise 4: Given the population mean is 80, and population standard deviation is 13.6. Suppose the sample size is 45.
a) What is the approximate probability that the sample average will be at least 115?

P(z ≥ #)

$z = \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \rightarrow \dfrac{115 - 80}{\frac{13.6}{\sqrt{45}}} = 17.27$ ↗ P(z ≥ 17.27) = 1 - P(z ≤ 17.27) = $\boxed{0}$

sampling distribution of a sample proportion

converting p → z : $z = \dfrac{\hat{p} - p}{\sigma_{\hat{p}}}$, $\hat{p} = \dfrac{\#\ \text{successes}}{n}$, $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$

p = 0.71

Exercise 7: During the holiday, it is estimated 71% of people travel to see family. If you survey 350 people n
a) What is the probability that less than 252 people say they travel?

$\frac{252}{350} = \hat{p}$

$z = \dfrac{\frac{252}{350} - 0.71}{\sqrt{\frac{0.71(1-0.71)}{350}}} = 0.42$ P(z < 0.42) = 0.6628

confidence interval

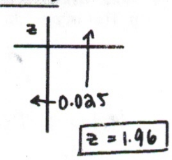
① $\left[\hat{p} - z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}\right]$

② σ is given - $\left[\bar{x} - z\dfrac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z\dfrac{\sigma}{n}\right]$ (use z!!)

③ s (standard dev.) is given $\left[\bar{x} - t\dfrac{s}{\sqrt{n}} \le \mu \le x + t\dfrac{s}{\sqrt{n}}\right]$ (t)

↳ $s = \sqrt{\dfrac{\Sigma(x - \bar{x})^2}{n-1}}$

finding z-score (ex: 95% conf)
① 100 - 95 = 5
② $\frac{5}{2}$ = 2.5 = 0.025
③ find corresponding z-value



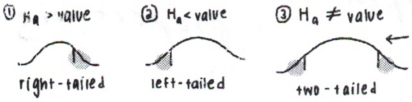
$\boxed{z = 1.96}$

90% → z = 1.64
95% → z = 1.96
99% → z = 2.57

finding t-score (ex: 95 conf) n = 43
① 100 - 95 = 5%
② $\frac{5}{2}$ = 0.025
③ find degrees of freedom → df = n-1
↳ df = 43 - 1 = 42
④ find t = 0.025 @ df = 42
$\boxed{t = 2.0}$

· increasing sample size = ↓ confidence interval
· decreasing sample size = ↑ confidence interval

① H_a > value ② H_a < value ③ H_a ≠ value



find area (p-value) of one side and multiply by 2

reject H_0 := P-value ≤ a
fail to reject H_0 = P-value > a

$z = \dfrac{\hat{p} - p}{\sigma_{\hat{p}}} = \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ ★

Ex: Exercise 4: According to a survey of 1000 adult Americans, 210 of those surveyed said playing the lottery would be the most practical way for them to accumulate $200,000 in net wealth in their lifetime. Although the article does not describe how the sample was selected, for purposes of this exercise, assume that the sample can be regarded as a random sample of adult Americans. Is there convincing evidence that more than 20% of adult Americans believe that playing the lottery is the best strategy for accumulating $200,000 in net wealth? Use α = 0.05.

H_a = p > 0.2 (right - tailed)
H_0 = 0.2

n = 1,000 $\hat{p} = \frac{210}{1,000}$ a = 0.05

$z = \dfrac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \dfrac{0.21 - 0.2}{\sqrt{\frac{(0.2)(1-0.2)}{1,000}}} = 0.79$

P(z > 0.79) = 1 - 0.7852 = 0.2148

0.2148 > 0.05 → fail to reject H_0 (no convincing evidence that H_a = p > 0.2)

if population mean is given...

$t = \dfrac{\bar{x} - \mu}{s_{\bar{x}}} = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$ df = n - 1 ★

$z = \dfrac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$

Ex: Exercise 8: When people smoke, the nicotine they absorb is converted to cotinine, which can be measured. A sample of 25 smokers has a mean cotinine level of 182.5 ng ml. Assuming that the sample standard deviation is known to be 114.5 ng ml, test the claim that the mean cotinine level of all smokers is equal to 200 ng ml. Use a 0.05 significance level.

n = 25 $\bar{x}$ = 182.5 ml s = 114.5 ng

H_0 = 200 , H_A = p ≠ 200 (two tailed)

$t = \dfrac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \dfrac{182.5 - 200}{\frac{114.5}{\sqrt{25}}} = -0.8$ @ df = 24

(0.216)(2) = 0.432 > 0.05 (fail to reject H_0)
(0.216: p - value; 2: two - tailed)

① comparing independent (standard deviation)

$V_1 = \dfrac{(s_1)^2}{n_1}$ $V_2 = \dfrac{(s_2)^2}{n_2}$

$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{V_1 + V_2}}$ ($\mu_1 - \mu_2$ ← 0)

$df = \dfrac{(V_1 + V_2)^2}{\dfrac{(V_1)^2}{n_1 - 1} + \dfrac{(V_2)^2}{n_2 - 1}}$

$s = \sqrt{\dfrac{\Sigma(x - \bar{x})^2}{n-1}}$

o: $\mu_1 - \mu_2$ = hyp. val. (0) Ha: $\mu_1 > \mu_2$

Exercise 2: The paper investigated the driving behavior of teenagers by observing their vehicles as they left a high school parking lot and then again at a site approximately mile from the school. Assume that it is reasonable to regard the teen drivers in this study as representative of the population of teen drivers.

Data consistent with summary quantities appearing in the paper are given in the table. The measurements represent the difference between the observed vehicle speed and the posted speed limit (in miles per hour) for a sample of male teenage drivers and a sample of female teenage drivers. Do these data provide convincing support for the claim that, on average, male teenage drivers exceed the speed limit by more than do female teenage drivers? Use a 0.01 level of significance.

Amount by Which Speed Limit Was Exceeded

Male Driver	Female Driver
1.2	0.3
1.4	[illegible]
[illegible]	1.2
[illegible]	0.7
[illegible]	1.1
[illegible]	1.2
[illegible]	0.1
1.3	0.9
0.8	0.8
[illegible]	0.8

male = 1 , female = 2 $\mu_1 - \mu_2 > 0$ (right tailed)

1 = male , 2 = female

n_1 = 10	n_2 = 10
$\bar{x}_1$ = 1.46	$\bar{x}_2$ = 0.64
V_1 = 0.055	V_2 = 0.019

$s(\text{male}) = \sqrt{\dfrac{(1.2 - 1.46)^2 + (1.4 - 1.46)^2 \ldots}{10 - 1}} = 0.741$

s(female) = 0.440

$t = \dfrac{(1.46 - 0.64) - (0)}{\sqrt{0.055 + 0.019}} = 3.0$ $df = \dfrac{(0.055 + 0.019)^2}{\frac{(0.055)^2}{9} + \frac{(0.019)^2}{9}} = 14$

t value @ df = 14 → 0.004 < 0.01 reject H_0 , support H_A

② comparing two dependent events (where $n_1 = n_2$)

$$t = \frac{\bar{x}_d - (\mu_1 - \mu_2)^{\to 0}}{\frac{s_d}{\sqrt{n}}} \qquad \bar{x}_d = \frac{\Sigma \text{ diff}}{n} = \frac{(x_1 - y_1) + \dots}{n}$$

$d = n - 1$

each (x-y) is squared

whole thing is squared

$$s_d = \sqrt{\frac{\Sigma(\text{diff})^2 - \frac{(\Sigma \text{diff})^2}{n}}{n-1}}$$

Exercise 5: To determine if chocolate milk was as effective as other carbohydrate replacement drinks, nine male cyclists performed an intense workout followed by a drink and a rest period. At the end of the rest period, each cyclist performed an endurance trial where he exercised until exhausted and time to exhaustion was measure. Each cyclist completed the entire regimen on two different days. On one day the drink provided was chocolate milk and on the other day the drink provided was a carbohydrate replacement drink. Data consistent with summary quantities appear in the table below.

Cyclist	Time to Exhaustion (minutes)								
	1	2	3	4	5	6	7	8	9
Chocolate Milk	35.56	47.12	34.63	37.79	55.97	31.16	33.23	21.67	36.48
Carbohydrate Replacement	23.78	44.00	12.74	37.63	31.56	12.19	14.33	0.53	31.46

Is there sufficient evidence to suggest that the mean time to exhaustion is greater after chocolate milk than after carbohydrate replacement drink? Use a significance level of 0.05.

$$\bar{x} = \frac{\Sigma \text{ diff}}{n} = \frac{122.43}{9} \qquad s_d = \sqrt{\frac{\Sigma(\text{diff})^2 - \frac{(\Sigma \text{diff})^2}{n}}{n-1}} = \sqrt{\frac{2244.5 - \frac{(14989.1)}{9}}{8}} = 8.508$$

$$t = \frac{\bar{x}_d - (\mu_1 - \mu_2)}{\frac{s_d}{\sqrt{n}}} = \frac{13.603 - 0}{\frac{8.508}{\sqrt{9}}} = 4.8 \quad @ \text{ df} = 8$$

df / t	8
4.8	0.001

p-value ≤ α

0.001 < 0.05

reject H_0, support H_a

③ comparing two independent events involving proportion

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_c(1-\hat{p}_c)}{n_1} + \frac{\hat{p}_c(1-\hat{p}_c)}{n_2}}} \qquad \hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Exercise 8: An article claimed that "those with a college degree reported a higher incidence of sunburn than those without a high school degree – 43 percent versus 35 percent." For purposes of this exercise, suppose that these percentages were based on random samples of size 200 from each of the two groups of interest. Is there convincing evidence that the proportion who experience a sunburn is higher for college graduates than it is for those without a high school degree? Use 0.05 significance level.

1 = college graduate   2 = w/o h.s degree

$n_1 = 200 \qquad n_2 = 200$

$\hat{p}_1 = 0.43 \qquad \hat{p}_2 = 0.35$

$p_1 > p_2$

$p_1 - p_2 > 0$

$$\hat{p}_c = \frac{200(0.43) + 200(0.35)}{200 + 200} = 0.39$$

$$z = \frac{0.43 - 0.35}{\sqrt{\frac{0.39(1-0.39)}{200} + \frac{0.39(1-0.39)}{200}}} = 1.64$$

z-1.64

$P(z > 1.64) = 1 - 0.9495$

$= 0.0505 > 0.05$

failed to reject H_0, failed to support H_a

---

chi-square (frequency):

① determine H_0 and H_a

② find sample size and proportion

③ calculate chi-square test statistic

④ find degree of freedom

⑤ find p-value from chi-square table

⑥ compare with α (same rules)

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \overbrace{\text{expected cell count}}^{n(p)})^2}{\text{expected cell count}}$$

$df = k - 1$

↳ number of categories

Exercise 3: A particular report included in the following table classifying 715 fatal bicycle accidents according to time of day the accident occurred.

Time of Day	Number of Accidents
Midnight to 3 A.M.	37 → expected = $(715)(\frac{1}{8})$ = 89.375
3 A.M. to 6 A.M.	28 → expected = $(715)(\frac{1}{8})$ = 89.375
6 A.M. to 9 A.M.	67
9 A.M. to Noon	77
Noon to 3 P.M.	99
3 P.M. to 6 P.M.	127
6 P.M. to 9 P.M.	166
9 P.M. to Midnight	114

Assume it is reasonable to regard the 715 bicycle accidents summarized in the table as a random sample of fatal bicycle accidents in that year. Do these data support the hypothesis that fatal bicycle accidents are not equally likely to occur in each of the 3-hour time periods used to construct the table? Test the relevant hypotheses using a significance level of 0.05. H_0 = equally H_a = not equally

Note: When p is not stated in the questions, we can assume each category shares the same p. That is, p = 1/k, and k = number of category.

↳ $\frac{1}{8}$ for all p

$$X^2 = \frac{(37 - 89.375)^2}{89.375} + \frac{(28 - 89.375)^2}{89.375} \dots = 169.509$$

expected preportions are given / assume

indepent variable preportions

X^2	df = 7
<0.001	>24.32

p-value > α

>0.1 >0.05

failed to reject H_0

chi-square (two way table):

① determine H_0 and H_a

② find expected value for each category

③ find X^2

④ find degree of freedom

⑤ find p-value and compare with α

$$\text{expected cell count} = \frac{(\text{row total})(\text{column total})}{\text{grand total}}$$

df = (number of rows − 1)(number of columns − 1)

Exercise 6: A particular paper described a study of children who were underweight or normal weight at age 2. Children in the sample were classified according to the number of sweet drinks consumed per day and whether or not the child was overweight one year after the study began. Is there evidence of an association between whether or not children are overweight after one year and the number of sweet drinks consumed? Assume that it is reasonable to regard the sample of children in this study as representative of 2 to 3 years old children and then test the appropriate hypotheses using a 0.05 significance level.

$\frac{(952)(253)}{8,328} = 28.9$ (expected)

$\frac{(952)(8,075)}{8,328} = 923.1$ (expected)

Number of Sweet Drinks Consumed per Day	Overweight?		
	Yes	No	total
0	22	930	952
1	73	2074	2147
2	56	1681	1,737
3 or More	102	3390	3492
	253	8,075	8,328

$$X^2 = \frac{(22-28.9)^2}{28.9} + \frac{(930-923.1)^2}{923.1} + \frac{(73-65.2)^2}{65.2} + \frac{(2074-2081.8)^2}{2081.8} + \frac{(56-52.8)^2}{52.8} + \frac{(1681-1684.2)^2}{1684.2} + \frac{(102-106.1)^2}{106.1} + \frac{(3390-3385.9)^2}{3385.9}$$

$= 3.025$ @ df = (4−1)(2−1) = 3

X^2	df = 3
0.1	<6.25

p-value > α

>0.1 > 0.05

fail to reject H_0