JOINTLY DISTRIBUTED RANDOM VARIABLES

Joint probability: What is P(X=2.5 and Y=0) → Cell in table where these two criteria meet=0.03
Marginal Distribution: Calculate marginal probabilities P(X) and P(Y) → Sum of each row and column
Test for independence: Are X and Y independent → Check if P(X and Y) = P(X)*P(Y) for each cell.

	Happiness (Y)			
Salary (X)	0	1	2	$P_X(x)$
2.5	.03	.12	.07	.22
7.5	.02	.13	.11	.26
12.5	.01	.13	.14	.28
17.5	.01	.09	.14	.24
$P_Y(y)$	.07	.47	.46	1.0

Conditional Distribution: Cond. distr. of salary given very happy P(X=x|Y=2)
Unconditional Distribution(Used to calculate expected values): P(X=x)
Test for independence: Conditional distr. ≠ Unconditional distr.

x	P(X=x|Y=2)
2.50	.07/.26=.15
7.50	.11/.46=.24
12.50	.14/.46=.305
17.50	.14/.46=.305

Conditional distribution of salary

x	P(X=x)
2.50	0.22
7.50	0.26
12.50	0.28
17.50	0.24

Unconditional distribution of salary

Conditional Expectation: Given Y is some value, what is expected value X
Test for independence: If X and Y were independent, conditional expectation would be same as overall expectation

$$E(X|Y=y) = \sum_{\text{all } x \text{ values}} xP(X=x|Y=y)$$

E[X|Y=0]=2.5(.03/.07)+7.5(.02/.07)+12.5(.01/.07)+17.5(.01/.07) = 7.45

COMBINING RANDOM VARIABLES

X and Y are independent: E[X+Y] = E[X]+E[Y] Var(X+Y)=VAR(X)+VAR(Y)
X and Y are dependent: E[X+Y] = E[X]+E[Y] Var(X+Y) = Var(X) + Var(Y)+2Cov(X,Y)
General: E[(a+bX)(c+dY)] = a+bE[X]+c+dE[Y] Var = b^2Var(X) + d^2Var(Y)+2bdCov(X,Y)
General: If X,Y are normally distributed, the sum ax+by is normal distributed.
NOTE: σ_x^2 = Var(X) = E[X^2] - μ^2

COMBINING FOR COMPARISON

Suppose two rats A and B have been trained to navigate a large maze.
X=Time of run for ratA X~N(80,10^2) Y=Time of run for rat B Y~N(78,13^2)
On any given day what is the probability that rat A runs the maze faster than rat B?

$$\begin{aligned} P(D<0) &= P\left(\frac{D-2}{\sqrt{269}} < \frac{0-2}{\sqrt{269}}\right) \\ &= P(Z<-0.122) \\ &= 0.4514 \end{aligned}$$

For a RV D = X-Y → E[D] = E[X-Y] and Var(D) = Var(X+Y) → P(X<Y) = P(D<0)

COVARIANCE

Measure of linear association of 2 RV. Sign reflects direction of association + → Same direction — → Opposite
Calculation:
a) COV(X,Y) = E[XY]-E[X]E[Y] b) $\sigma_{XY} = \sum_{i=1}^{N}[x_i - E(X)][(y_i - E(Y)]\,P(x_i, y_i)$

Example: Cov(X,Y) = 3.89 - (2.59)(1.52) = -0.05

	x				
y	1	2	3	4	Total
1	.04	.14	.23	.07	.48
2	.07	.17	.23	.05	.52
Total	.11	.31	.46	.12	1

$$\begin{aligned} E(XY) &= \sum_{j=1}^{m}\sum_{i=1}^{n}(X_iY_j)P(X_iY_j) \\ &= (1)(1)(.04) + (1)(2)(.14) + (1)(3)(.23) + (1)(4)(.07) + (2)(1)(.07) \\ &\quad + (2)(2)(.17) + (2)(3)(.23) + (2)(4)(.05) \\ &= 3.89 \end{aligned}$$

$$\mu_X = \sum_{i=1}^{4} X_iP(X_i) = 1(.11) + 2(.31) + 3(.46) + 4(.12) = 2.59$$

$$\mu_Y = \sum_{j=1}^{2} Y_jP(Y_j) = 1(.48) + 2(.52) = 1.52$$

Properties: 1) Cov(X,X) = Var(X) 2) If X,Y indep. Cov(X,Y) = 0 3) Depends on units of measure
Uses: Lowest covariance in combination of stocks is lowest risk

CORRELATION

Dimensionless measure.
Always between -1 and 1.
-1 → perfectly negative(zero risk portfolio) +1 → perfectly positive
Test for independence: Correlation is 0 → no linear relationship
Uses: If two securities are negatively correlated, it would reduce total risk.

$$\rho = \frac{\sigma_{X,Y}}{\sigma_X\sigma_Y} = \frac{-2.91}{\sqrt{2.91}\sqrt{2.91}} = -1$$

CENTRAL LIMIT THEOREM

CLT let's you answer questions about X' i.e. sample mean.
We can't answer questions about the population based on mean and variance when we don't know distribution.

FOR CONTINUOUS DATA

1. If samples of size n ≥ 30 are drawn from any population with mean = μ and standard deviation = σ, then the sampling distribution of the sample means approximates a normal distribution. The greater the sample size, the better the approximation.

2. If the population is normally distributed, then the sampling distribution of sample means is normally distributed for any sample size n (not just ≥ 30).

Sample mean normally distributed as: $X' \sim N(\mu, \frac{\sigma^2}{n})$ Standard Deviation or Standard error of sample mean: $\sigma_X = \frac{\sigma}{\sqrt{n}}$

Unless the distribution is explicitly told, it is not possible to evaluate P (a < X < b). However, with n sufficiently large, the CLT allows one to evaluate P (a < X < b) irrespective of the underlying population.

$$P(7.8 < \bar{X} < 8.2) = P\left(\frac{7.8-8}{3/\sqrt{36}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{8.2-8}{3/\sqrt{36}}\right) = P(-0.4 < Z < 0.4) = 0.3108$$

Examples:
1) What is the probability that the sample mean will be within 300 miles of the population mean? $P(\mu - 300 < \bar{X} < \mu + 300)$
2) The service times for customers coming through a checkout counter in a retail store are independent random variables with a mean of 1.5 minutes and a variance of 1.0. Approximate the probability that 88 customers can be serviced in less than 2 hours of total service time by this one checkout counter. $P(\sum X_i < 120) = P(\bar{X} < 120/88)$ with $\bar{X} \sim N(1.5, 1/88)$

FOR A PROPORTION

For discrete data, we want to know the proportion of a set of values. Assume p is the population proportion of a given characteristic. We now want to estimate p using a sample of the population. This estimate is called $\hat{p}$.

$\hat{p} = \frac{x}{n} = \frac{number of successes in sample}{sample size}$ $E[\hat{p}] = p$ $Var(\hat{p}) = \frac{pq}{n}$ Std Dev = $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

$\hat{p} \sim N(n, \frac{p(1-p)}{n})$

For discrete data n * p ≥ 5 and n(1−p) ≥ 5 for the CLT to "kick in." In practice n has to relatively much larger like > 100.

BIAS OF AN ESTIMATOR

Guesses should be unbiased and have minimum variation(MVUE)

bias($\hat{\theta}$) = E[$\hat{\theta}$] - θ where $\hat{\theta}$ is estimate of some param θ. An estimate is unbiased if its bias = 0

CONFIDENCE INTERVAL(USED FOR TWO SIDED ALTERNATIVE)

95% confidence interval means that 95% of samples of this size will produce confidence intervals that capture the true proportion, or we are 95% confident that the true proportion lies in our interval.

Confidence Level	Confidence Coefficient, $1-\alpha$	z value $Z_{\alpha/2}$
80%	.80	1.28
90%	.90	1.645
95%	.95	1.96
98%	.98	2.33
99%	.99	2.58
99.8%	.998	3.08
99.9%	.999	3.27

CONFIDENCE INTERVAL OF A MEAN

$\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$ where margin of error: $\pm 1.96\frac{s}{\sqrt{n}}$

For confidence other than 95% At 95%, $\alpha = 5\% = 0.05$ $\alpha/2 = 0.025$ $Z_{\alpha/2} = 1.96$

Small Sample(data needs to be normally distributed): If n < 30, we need σ not s and we use t-distribution i.e. $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$

Replace 1.96 with the t-value instead of Z-value
NOTE: In the t-value sheet, df. Row 1 is used for n = 2 and so on

For 90% ci's → $\alpha/2$ ← For 99% ci's

d. f.	0.100	0.050	0.025	0.010	0.005
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925

So if n=2 and you want to compute a 95% confidence interval (you doofus), it would be:

$$\bar{x} \pm 12.706\left(\frac{s}{\sqrt{n}}\right)$$

But if n=29 and you want to compute a 95% confidence interval it would be:

$$\bar{x} \pm 2.048\left(\frac{s}{\sqrt{n}}\right)$$

CONFIDENCE INTERVAL OF A PROPORTION

$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = \hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ where margin of error: $\pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

Sample size: $n = \frac{(1.96)^2\hat{p}(1-\hat{p})}{e^2}$ for a e % margin of error

How to determine $\hat{p}$: 1) From pilot study 2) Use 0.5 for worst case estimate

HYPOTHESIS TESTING(USED FOR ONE SIDED ALTERNATIVE)

Null Hypothesis: H_0 is a statement to be tested. The null hypothesis is a statement of no change, no effect or no difference (status quo). It is assumed true until evidence indicates otherwise. By default, we assume that nothing has changed, and then try to disprove it. It always has an equal sign.
Alternative Hypothesis: H_a is the alternative hypothesis, which we are trying to find evidence to support. Since we only have sample data, we can really only disprove a theory, not prove it, since we haven't seen all the data. (Basically we are trying to find the exception to the H_0 null hypothesis). It is always easier to disprove than to prove something.
Two tail test: H_0 = some value; H_a/= some value Left tail test: H_0 = some value; H_a < some value
Right tail test: H_0 = some value; H_a > some value

HYPOTHESIS TESTING OUTCOMES

Correct decisions: Reject null hypothesis when alternative is correct. Do not reject null hypothesis when alternative is correct.
Type I error: Reject null when null is correct
Type II error: Do not reject null when alternative is correct

Usually, you can minimize either Type I or Type II errors, but not both due to an inverse relationship. It is worse to make Type I errors, so that is usually minimized.

Level of significance: α = P (Type I error) = P(rejecting H_0 when it is true) = P (reject H_0|H_0 is true)
We never "accept" the null hypothesis because we don't have access to the entire population. Instead we don't reject the null hypothesis.

β = P(Type II error) = P (not rejecting H_0 when H_a is true)

TWO TAILED TEST USING CONFIDENCE INTERVAL

1. Define Hypothesis $H_0 : \theta = \theta_0$ $H_a : \theta \neq \theta_0$
2. Construct Confidence Interval for mean or proportion
3. Accept or Reject: If θ_0 falls within this interval, we fail to reject the null. If θ_0 is outside the interval, we reject the null.

TEST STATISTIC APPROACH

Assuming n > 30, test for t_{stat}
We want to check if μ is near μ_0

$$t_{stat} = \frac{\bar{x} - \mu_o}{s / \sqrt{n}}$$

If it is, we fail to reject null hypothesis.
Otherwise, reject.

For n < 30, 1.96 and 1.64 need to be adjusted using t-values. So use p values instead

The 1.96 and 1.64 are at a 5% significance level.

Test Statistic	Decision Rule
$H_0 : \mu = \mu_0$ $H_a : \mu \neq \mu_0$	If $\lvert t_{stat} \rvert > 1.96$, reject H_0
$H_0 : \mu \geq \mu_0$ $H_a : \mu < \mu_0$	If $t_{stat} < -1.64$, reject H_0
$H_0 : \mu \leq \mu_0$ $H_a : \mu > \mu_0$	If $t_{stat} > 1.64$, reject H_0

TESTING A PROPORTION

Test Statistic	Decision Rule
$H_0 : p = p_0$ $H_a : p \neq p_0$	If $\lvert t_{stat} \rvert > 1.96$, reject H_0
$H_0 : p = p_0$ $H_a : p < p_0$	If $t_{stat} < -1.64$, reject H_0
$H_0 : p = p_0$ $H_a : p > p_0$	If $t_{stat} > 1.64$, reject H_0

$$t_{stat} = \frac{(\hat{p} - p_0)}{\sqrt{p_0(1-p_0)/n}} \quad \text{or} \quad t_{stat} = \frac{(\hat{p} - p_0)}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

For proportions we always assume a lot of data $n \gg 30$.

P - VALUES

Probability Values (P-values) in range [0,1] that provide strength of evidence. If P is low, H_0 must go. It is a measure of how much statistical evidence exists.

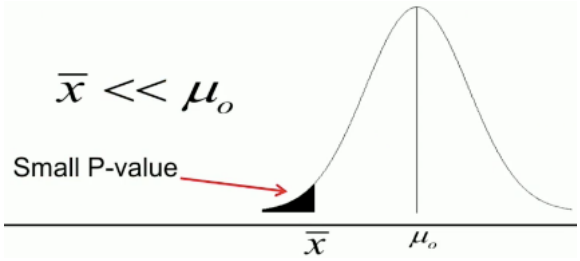
If P is high, there is evidence for H_0.
If P is low, there is evidence for H_a.

It is a measure of how consistent the data is with the null hypothesis.

p-value < 0.01 corresponds to "highly statistically significant, very strong evidence" to reject H_0.
0.01 < p-value < 0.05 corresponds to "statistically significant and adequate evidence" to reject H_0.
p-value > 0.05 corresponds to "insufficient evidence" against H_0.

Good things about p-values:
• automatically adjusts for large and small datasets
• don't have to worry about one-sided or two-sided
• just simply read the value and plug into the appropriate range

The P-value is the Cumulative Distribution Function (the shaded area under the Gaussian distribution curve).
The smaller it is, the further x is from μ_0.
Therefore, p < 0.05, means x is so far from μ_0 that we must reject H_0
If x = μ_0 then the P-value = 0.5 (half of area under normal distribution would be filled in).



Need to calculate:
To find p-value for a sample X', we use Z-score
P (x < value) where x~N(μ_0,$\frac{\sigma^2}{n}$)

Example: Calculate p-value for X' = 27.80 given
μ_0 = 30.5, σ = 6.6 and n = 36

We need to calculate

$$P(\bar{X} < 27.80) \quad where\ \bar{X} \sim N(30.5, 1.21)$$

Using the Z score

$$P(\bar{X} < 27.80) \quad where\ \bar{X} \sim N(30.5, 1.21)$$
$$= P\left(\frac{\bar{X} - 30.5}{\sqrt{1.21}} < \frac{27.80 - 30.5}{\sqrt{1.21}}\right)$$
$$= P(Z < -2.45) = 0.0071$$

NOTE: This is testing X' < value i.e. Ha: μ < value. For testing Ha: μ > value, we flip the sign in Z-score calculation but 0.05 condition for p-value still stands.

COMPARING TWO GROUPS

CONFIDENCE INTERVAL OF DIFFERENCE OF TWO MEANS

$$(\bar{x} - \bar{y}) \pm 1.96(\sqrt{(\frac{Sx}{n2} + \frac{Sy}{n1})})$$

CONFIDENCE INTERVAL OF DIFFERENCE OF TWO PROPORTIONS

$$(\hat{p}1 - \hat{p}2) \pm 1.96\sqrt{(\frac{\hat{p}1(1-\hat{p}1)}{n1} - \frac{\hat{p}2(1-\hat{p}2)}{n2})}$$

If the interval is all positive then $\hat{p}_1 > \hat{p}_2$.

If the interval is all negative then $\hat{p}_1 < \hat{p}_2$.

If the interval spans 0, then one is not significantly bigger than the other (or cannot be determined).

As long as $n > 30$, it doesn't matter if the sample size is different between the random variables.

Interpretation: The true difference in proportion is in the confidence interval and which is greater is determined by above logic.
Example: Confidence interval 0.16 - 0.27 → Proportion 1 > Proportion 2 and the decrease is between 16 and 27%.
Assumption: To two difference of proportions, we need to assume the two proportions are independent.

TESTING TWO PROPORTIONS

Decision Rules for Testing Two Proportions

$$T = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}, \text{where } \hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

$H_0 : p_1 = p_2$ $H_a : p_1 \neq p_2$	If $\lvert T \rvert > 1.96$ *reject* H_o
$H_0 : p_1 = p_2$ $H_a : p_1 < p_2$	If $T < -1.64$ *reject* H_o
$H_0 : p_1 = p_2$ $H_a : p_1 > p_2$	If $T > 1.64$ *reject* H_o

TESTING TWO MEANS

Decision Rules for Testing Two Samples

$$T = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$ the test statistic

$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 \neq \mu_2$	If $\lvert T \rvert > 1.96$ *reject* H_o
$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 < \mu_2$	If $T < -1.64$ *reject* H_o
$H_0 : \mu_1 = \mu_2$ $H_a : \mu_1 > \mu_2$	If $T > 1.64$ *reject* H_o Assuming both sample sizes > 30

MATCHED PAIRS (THE TWO SAMPLES ARE NOT INDEPENDENT)

Example: Before and after samples of Weight-watchers group
Can not rely on p-value for the two samples here
Create a new sample of the difference of the two samples and do a one sided test on the difference.

$$H_0 : \mu_{new} = \mu_{old} \qquad H_a : \mu_{new} > \mu_{old}$$

Define the difference

$$D = NewScore - OldScore$$

We want to test

$$H_0 : \mu_D = 0 \qquad H_a : \mu_D > 0$$

CHI-SQUARED GOODNESS OF FIT

Tests several proportions at the same time, aka the multinomial setting.
k categories of interest with $p_1, p_2, \ldots, p_k$ probabilities that a value is in a particular cell.
All p's add up to 1, as usual.

$H_0 : p_1 = a_1, p_2 = a_2, \ldots, p_k = a_k$
where $a_1, a_2, \ldots, a_k$ are the values to be tested.
H_a : at least one p_i is not equal to the specified value.
O observed frequency of an outcome, given
E expected frequency of an outcome, calculated
k number of different categories

n number of trials
s_i sample standard deviation

Calculate Observed and Expected to see if they are consistent. Known as Chi-Squared Goodness of Fit (GOF) Test.
$e_i = n \cdot p_i$

Smallest possible value is zero. Smaller χ^2 means H_0 is plausible. Larger χ^2 means reject the null.

$$\chi^2 = \sum_{\text{all } i} \frac{(o_i - e_i)^2}{e_i}$$

Use table to determine cut-off values (determined by degrees of freedom k − 1). As before, we typically use α = 5% level of significance.

If $\chi^2 > \chi^2_{\alpha,k-1}$, then reject the null in favor of H_a. Something has changed (but we don't know what or which direction). $\chi^2_{\alpha,k-1}$ comes from Chi-squared table for column α and row(df) k-1 so 3-1 = 2 for example to the right

Example:

Requirements:
1. Data is random
2. Data has frequency counts per category
3. $e_i \geq 5$, o_i can be anything. Might need to group smaller categories.

In order to calculate our test statistic, we lay-out the data in a tabular fashion for easier calculation by hand:

Company	Observed Frequency	Expected Frequency	Delta	Summation Component
	o_i	e_i	$(o_i - e_i)$	$(o_i - e_i)^2/e_i$
A	102	90	12	1.60
B	82	80	2	0.05
Others	16	30	-14	6.53
Total	200	200		**8.18**

Check that these are equal

$$\chi^2 = \sum \frac{(o_i - e_i)^2}{e_i}$$

CHI-SQUARED TEST OF INDEPENDENCE

aka Two-way Chi-Squared Test.
Tests if r rows and c columns are independent or not. H_0 is independent, H_a is depen- dent.
Need to figure out the probabilities in order to determine e_i.
Recall, for independent variables:
P(A and B) = P(B)P(A)
If $e_{ij} = P(r_i)P(c_j)$ then independent

Our rejection region is:

$$\chi^2 > \chi^2_{\alpha,k-1} = \chi^2_{.05,3-1} = 5.99147$$

Since our test statistic is 8.18 which is greater than our critical value for Chi-squared, we reject H_0 in favor of H_a, that is,

"There is sufficient evidence to infer that the proportions have changed since the advertising campaigns were implemented"

ANOVA

One Way ANOVA
The one-way analysis of variance method is used to test the claim that three or more population means are equal
This is an extension of the two independent sample t-test
Its called analysis of variance because it works by (non intuitively) comparing different sample variances.
Null hypothesis is that all means are equal.

Why not several 2 sample tests?
First, when you are comparing two means at a time, the rest of the means under study are ignored. With ANOVA, all the means are compared simultaneously.
Second, the more means there are to compare, the more t tests are needed. For example, for the comparison of 3 means two at a time, 3 t tests are required. For the comparison of 5 means two at a time, 10 tests are required. And for the comparison of 10 means two at a time, 45 tests are required.

Terminology:
The *response* variable is the variable you're comparing The *factor* variable is the categorical variable being used to define the groups
We will assume *k* samples (groups)
The *one-way* is because each value is classified in exactly one way
Examples include comparisons by gender, race, political party, color, etc

Guiding principle:
ANOVA works by **assuming** all the data has the same variance. The only possible difference is with the means
It turns out there are two different ways to estimate this common variance; so we do that and compare the two values.