Probability

P(A or B): $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

For any two events, P(A and B) = $P(A \cap B) = P(A|B)P(B) = P(B|A)\,P(A)$

P (A and B and C)

$P(A \cap B \cap C) = P(A \cap B)P(C|A \cap B) = P(A)P(B|A)\,P(C|A \cap B)$,

$P(A \cap B \cap C \cap D) = P(A)\,P(B|A)P(C|A \cap B)\,P(D|A \cap B \cap C)$,

Two events are independent if $P(A|B) = P(A)$ or $P(B|A) = P(B)$;

For independent events A and B, $P(A \cap B) = P(A)\,P(B)$,

$P(A|B) = \frac{P(A \cap B)}{P(B)}$

Marginal Probability Rule: $P(B) = \sum_{i=1}^{k} P(A_i \cap B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$

A and B are mutually exclusive events if and only if either of the following holds:
P(A|B) = 0 or P(B|A) = 0.

Bayes Theorem:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{i=1}^{k} P(B|A_i)P(A_i)}$$

P(R) = P(R|B1) P(B1) + P(R|B2) P(B2) + P(R|B3) P(B3)

General Formulae

$\sigma^2 = \frac{\sum(X-\overline{X})^2}{N}$ $\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(X-\overline{X})^2}{N}}$ $s^2 = \frac{\sum(X-\overline{X})^2}{N-1}$

$s = \sqrt{s^2} = \sqrt{\frac{\sum(X-\overline{X})^2}{N-1}}$ $\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{n-1}$ $r = \frac{\text{cov}(X,Y)}{S_X S_Y}$

Discrete Probability Distributions

Expected Value E(X): $\mu = E(X) = \sum_{i=1}^{N} X_i P(X_i)$ E(aX+b) = aE(x) + b

Variance (one variable): $\sigma^2 = \sum_{i=1}^{N}[X_i - E(X)]^2 P(X_i)$

$$\text{Var}(aX+b) = \text{Var}(aX) = a^2\text{Var}(X).$$

St.Dev (one variable): $\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{N}[X_i - E(X)]^2 P(X_i)}$

Covariance between two variables: Let X = one fund and Y = another fund

$$\sigma_{XY} = \sum_{i=1}^{N}[X_i - E(X)][(Y_i - E(Y)]\,P(X_i Y_i)$$

$$\text{Cov}(aX, bY) = ab\,\text{Cov}(X,Y)$$

Expected Value of the SUM of two random variables:

$$E(X+Y) = E(X) + E(Y)$$

Variance of the SUM of two random variables:

$$\text{Var}(X+Y) = \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\sigma_{XY}$$

$$\text{Var}(aX+bY) = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_{XY}$$

St. Dev of the SUM of two random variables: $\sigma_{X+Y} = \sqrt{\sigma^2_{X+Y}}$

Portfolio Expected Return (see units first):

$$E(P) = wE(X) + (1-w)E(Y)$$

Portfolio Expected Risk: $\sigma_P = \sqrt{w^2\sigma^2_X + (1-w)^2\sigma^2_Y + 2w(1-w)\sigma_{XY}}$

Binomial Distribution

The binomial distribution probability formula

$$P(X = x \mid n, \pi) = \frac{n!}{x!\,(n-x)!}\pi^x(1-\pi)^{n-x}$$

Where x = number of "events of interest" in sample, n = sample size, and π = probability of "event of interest"

Mean of the binomial Distribution: $\mu = E(X) = n\pi$

Variance of binomial distribution: $\sigma^2 = n\pi(1-\pi)$

St. Dev of Binomial Distribution: $\sigma = \sqrt{n\pi(1-\pi)}$

Poisson Distribution

The Poisson distribution probability: $P(X = x \mid \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$

Mean = λ (lambda) (avg no. of events p/unit) Variance = λ (lambda)

Standard Deviation = $\sigma = \sqrt{\lambda}$

Hypergeometric Distribution

$$P(X = x \mid n, N, A) = \frac{[{}_A C_X][{}_{N-A} C_{n-X}]}{{}_N C_n} = \frac{\binom{A}{X}\binom{N-A}{n-X}}{\binom{N}{n}}$$

Mean: $\mu = E(X) = \frac{nA}{N}$

n = sample size
N = popu. Size
A = no. of events of interest in pop.
x = no. of events of interest in sample

St. Dev : $\sigma = \sqrt{\frac{nA(N-A)}{N^2}} \cdot \sqrt{\frac{N-n}{N-1}}$

Normal Distribution

Symmetrical: Mean = Median
Interquartile Range (middle 50% of data) = 1.33 Stdev, Range = 6 stdev
Empirical rule: 1 stdev: 68.26%, 2 stdev: 95.44%, 3 stdev: 99.73%

Probability Density Function: $f(X) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{(X-\mu)}{\sigma}\right)^2}$

Find probability getting a value in a certain range (probability of an exact value is 0):

Transformational Formula (Z value): $Z = \frac{X-\mu}{\sigma}$

Find probability in Z value table (area less than the Z value)

Find value when given probability

Work backwards from probability e.g. 10% of download times: 0.1

Backwards formula to find X value: $X = \mu + Z\sigma$

Normal Probability Plot: *Plot Z Scores*

Uniform Distribution:

Probability Density Function of the whole thing: $\frac{1}{b-a}$ if $a \le X \le b$

where a is the min value of X and b is the max value of X, Mean: $\mu = \frac{a+b}{2}$

Standard Deviation: $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Probability of getting a value in a certain range:

Calculate area of rectangle: Base x Height or (y-x)/(b-a)

Exponential Distribution:

Mean = λ = the mean no. of arrivals per unit of time, X = any value of the continuous variable

Probability that an arrival time is less than some specified time X is:

$$P(\text{arrival time} < X) = 1 - e^{-\lambda X}$$

Standard deviation of time between arrivals = Mean time between arrivals = 1/ λ

Converting normal to binomial: The normal distribution can be used to approximate the binomial distribution if nπ ≥ 5 and n(1 – π) ≥ 5. Formula below:

$$Z = \frac{X-\mu}{\sigma} = \frac{X - n\pi}{\sqrt{n\pi(1-\pi)}}$$

Populations and Sampling Distributions

Pop mean: $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$ Pop Var: $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$ Pop stdev: $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}$

Regardless of how the data are distributed, at least (1 - 1/k2) x 100% of the values will fall within k standard deviations of the mean (for k > 1)

Standard error of the mean = standard deviation of all sample means $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

Standard error of the mean for finite populations: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$

How to determine the probability that a sample has a mean of below a certain number?

Convert to Z value as in normal distribution: $Z = \frac{(\overline{X} - \mu_{\overline{X}})}{\sigma_{\overline{X}}} = \frac{(\overline{X}-\mu)}{\frac{\sigma}{\sqrt{n}}}$

Get Z score then see table for probability: x% of all possible samples will have a mean of below the number

Working backwards to find the sample mean from probability (Determining an Interval Including a Fixed Proportion of the Sample Means):

Calculate lower and upper limits $\overline{X}_L = \mu + Z\frac{\sigma}{\sqrt{n}}$, $\overline{X}_U = \mu + Z\frac{\sigma}{\sqrt{n}}$

Sample proportion to estimate population of interest proportion(π):

$$p = \frac{X}{n} = \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

Confidence Interval:

Confidence Interval estimate for μ

(σ Known) : $\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ (σ Unknown): $\overline{X} \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$

(where tα/2 is the critical value of the t distribution with n -1 degrees of freedom and an area of α/2 in each tail)

Confidence Interval estimate for the population proportion (π):

$$p \pm Z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

Determining the Sample Size:

For the Mean: $n = \frac{Z_{\alpha/2}{}^2\,\sigma^2}{e^2}$ For the Proportion: $n = \frac{Z^2\pi(1-\pi)}{e^2}$

Hypothesis Testing:

Think: Two-tail or One-tail?

Z Test of Hypothesis for the Mean (σ Known): $Z_{STAT} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$

p-value: Probability of obtaining a test statistic equal to or more extreme than the observed sample value given H0 is true

If p-value < α , reject H0; If p-value ≥ α , do not reject H0

If the p-value is low then H0 must go

t Test of Hypothesis for the Mean (σ Unknown): $t_{STAT} = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$

Proportions: When both nπ and n(1-π) are at least 5, p can be approximated by a normal distribution with mean and standard deviation in terms of the no. of interest, X

$Z_{STAT} = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$ $Z_{STAT} = \frac{X-n\pi}{\sqrt{n\pi(1-\pi)}}$

The power of the test (1-B) is the probability of correctly rejecting a false H0

Calculating B Risk, if a = 0.05, n = 64, σ = 6

$$\text{cutoff} = \overline{X}_\alpha = \mu - Z_\alpha\frac{\sigma}{\sqrt{n}} = 52 - 1.645\frac{6}{\sqrt{64}} = 50.766$$

$$P(\overline{X} \ge 50.766 \mid \mu = 50) = P\left(Z \ge \frac{50.766-50}{6/\sqrt{64}}\right) = P(Z \ge 1.02) = 1.0 - 0.8461 = 0.1539$$

Two-Sample Tests

Comparing the Means of Two Independent Populations:

H0: μ1 = μ2

H1: μ1 ≠ μ2

Pooled-Variance Test (two tail or one tail):

1. Random Samples are independent selected from the 2 populations
2. Populations are normally distributed and have equal variances (assume)

$t_{STAT} = \frac{(\overline{X}_1-\overline{X}_2)-(\mu_1-\mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ Where, $S_p^2 = \frac{(n_1-1)S_1{}^2+(n_2-1)S_2{}^2}{(n_1-1)+(n_2-1)}$

**tSTAT has degrees of freedom = (n1 + n2 – 2)

The confidence interval estimate for the difference between two means:

$(\overline{X}_1-\overline{X}_2) \pm t_{\alpha/2}\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ Where tα/2 has d.f. = n1 + n2 – 2

Separate-Variance t Test:

1. Cannot assume the two populations have equal variances

Chi-Square Test

χ2 Test for the Difference Between 2 Proportions(onetail): $\chi^2_{STAT} = \sum_{all\,cells}\frac{(f_o-f_e)^2}{f_e}$

Decision Rule: If $\chi^2_{STAT} > \chi^2_\alpha$, reject H0, otherwise, do not reject H0

The average proportion is: $\overline{p} = \frac{X_1+X_2}{n_1+n_2} = \frac{X}{n}$ Degrees of freedom = (r-1)(c-1)

χ2 Test of Independence(onetail): Same Test Stat $f_e = \frac{\text{row total} \times \text{column total}}{n}$

Chi-Square Test for a Variance or Standard Deviation (two tail): $\chi^2_{STAT} = \frac{(n-1)S^2}{\sigma^2}$

Decision Rule: Reject if $\chi^2_{STAT} > \chi^2_{\alpha/2}$ or if $\chi^2_{STAT} < \chi^2_{1-\alpha/2}$, d.f. = n-1

Comparing the Means of Two Related Populations:

Paried t test (two tail test): H0: μD = 0 H1: μD ≠ 0

The ith paired difference is Di , where, Di = X1i - X2i

The point estimate for the paired difference population mean μD is DBar : $\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n}$

$S_D = \sqrt{\frac{\sum_{i=1}^{n}(D_i-\overline{D})^2}{n-1}}$ $t_{STAT} = \frac{\overline{D}-\mu_D}{\frac{S_D}{\sqrt{n}}}$, where tSTAT has n - 1 degrees of freedom!, two tail

The confidence interval for μD is $\overline{D} \pm t_{\alpha/2}\frac{S_D}{\sqrt{n}}$

Comparing the Proportions of Two Independent Populations:

Z Test for the difference between two proportions (two/one tail) H0: π1 = π2 H1: π1 ≠ π2

The point estimate for the difference is $p_1 - p_2$

$Z_{STAT} = \frac{(p_1-p_2)-(\pi_1-\pi_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ $\bar{p} = \frac{X_1+X_2}{n_1+n_2}$, $p_1 = \frac{X_1}{n_1}$, $p_2 = \frac{X_2}{n_2}$

X1 = no. Of items of interest in sample 1

X2 = no. of items of interest in sample 2

The confidence interval for π1 – π2 is:

$(p_1-p_2) \pm Z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}$

Comparing the Means of Two Independent Populations:

F Test for the Ratio of Two Variances (either one or two tail, usually two tail):

Assumption: Populations are normally distributed. Remember Larger variance on top!!!

$H_0: \sigma_1^2 = \sigma_2^2$	$H_0: \sigma_1^2 \le \sigma_2^2$
$H_1: \sigma_1^2 \ne \sigma_2^2$	$H_1: \sigma_1^2 > \sigma_2^2$

$F_{STAT} = \frac{S_1^2}{S_2^2}$

In the F table,

numerator degrees of freedom determine the column (n1 – 1)

denominator degrees of freedom determine the row (n2 – 1)

If F_{STAT} above critical value, reject null

Linear Regression

Simple Linear Regression Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ Prediction Line: $\hat{Y}_i = b_0 + b_1 X_i$

Least Squares Method: $b_1 = \frac{\text{Cov}(X,Y)}{S_X^2} = \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum(X-\overline{X})^2}$ $b_0 = \overline{Y} - b_1\overline{X}$

Total Variation = SST = SSR + SSE

$SST = \sum(Y_i-\overline{Y})^2$ explained: $SSR = \sum(\hat{Y}_i-\overline{Y})^2$ unexp: $SSE = \sum(Y_i-\hat{Y}_i)^2$

Coefficient of Determination: $R^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$ $0 \le R^2 \le 1$

Standard Error of Estimate: $s_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{n-2}}$

(standard deviation of the variation of observations around the regression line)

Assumptions: Linearity, Independence of Errors, Normality of Error, Equal Variance

Standard Error of the regression slope coefficient (b_1) is estimated by:

$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)S_X^2}} = \frac{s_\varepsilon}{\sqrt{\sum(X_i-\overline{X})^2}}$

t test for a population slope: Is there a linear relationship between X and Y? (two tail)

H0: β1 = 0 (no relationship);H1: β1 ≠ 0 (relationship exist) $t_{STAT} = \frac{b_1-\beta_1}{S_{b_1}}$ df: n-2

F Test for Significance (one tail): $F_{STAT} = \frac{SSR}{MSE}$ d.f. num: 1; denom: n-2 $MSE = \frac{SSE}{n-2}$

Confidence Interval Estimate for the Slope: $b_1 \pm t_{\alpha/2}S_{b_1}$ d.f. = n-2 *If 0 is incl, no signi.*

95% confident that the average impact on sales price is between () and () per sqft

Confidence interval estimate for the mean value of Y given a particular Xi: d.f. = n-2

Confidence interval for $\mu_{Y|X=X_0}$: $\hat{Y} \pm t_{\alpha/2}s_\varepsilon\sqrt{h_i}$ $h_i = \frac{1}{n} + \frac{(X_i-\overline{X})^2}{\sum(X_i-\overline{X})^2}$

Confidence interval estimate for an Individual value of Y given a particular Xi: d.f. = n-2

Confidence interval for $Y_{X=X_0}$: $\hat{Y} \pm t_{\alpha/2}s_\varepsilon\sqrt{1+h_i}$