Statistics Formula Sheet

Center, spread and outliers

Mean	$\bar{x} = \sum x_i / n$
Sample variance	$s^2 = \sum (x_i - \bar{x})^2/(n-1)$
Interquartile range	$IQR = Q_3 - Q_1$
Potential outlier fences	$Q_1 - 1.5\,IQR,\ Q_3 + 1.5\,IQR$

Divide by N for the variance of a whole population. Sample SD is sqrt(s^2).

Median and IQR are less sensitive to extreme observations.

Normal model and z-scores

Standardize	$z = (x - \mu)/\sigma$
Convert back	$x = \mu + z\,\sigma$
Within 1, 2, 3 SDs	about 68%, 95%, 99.7%

Use positive sigma. The percentages describe a normal distribution.

Sampling distributions

Sample mean	$E(\bar{x}) = \mu,\ SD = \sigma/\sqrt{n}$
Sample proportion	$E(\hat{p}) = p,\ SD = \sqrt{p(1-p)/n}$
Standard error	

replace unknowns with estimates

The SD formulas assume independent observations. Check the 10% rule for samples without replacement.

Check approximate normality before using a normal critical value.

Events and their complements

Complement	$P(\text{not } A) = 1 - P(A)$
A or B	$P(A) + P(B) - P(A \text{ and } B)$
Disjoint events	$P(A \text{ and } B) = 0$
Independent events	$P(A \text{ and } B) = P(A)P(B)$

Disjoint events with positive probabilities cannot be independent.

Binomial and geometric trials

Binomial: k successes	$P(X = k) = C(n,k)p^k(1-p)^{n-k}$
Binomial moments	$E(X) = np,\ Var(X) = np(1-p)$
First success at trial k	$P(X = k) = (1-p)^{k-1}p$
Geometric mean	$E(X) = 1/p$

Use independent trials with constant success probability p. Geometric k starts at 1.

Confidence intervals for means

One mean	$\bar{x} \pm t^* s/\sqrt{n}$
Paired mean difference	$\bar{d} \pm t^* s_d/\sqrt{n}$
Independent groups	$\text{estimate} = \bar{x}_1 - \bar{x}_2$
Independent group SE	$\sqrt{s_1^2/n_1 + s_2^2/n_2}$

Use t with appropriate degrees of freedom. For paired data analyze the differences.

Check randomization, independence and the shape or sample size of each sample.

Confidence intervals for proportions

One proportion	$\hat{p} \pm z^* \sqrt{\hat{p}(1-\hat{p})/n}$
Two proportions	$\text{estimate} = \hat{p}_1 - \hat{p}_2$
Two proportions	$\text{interval} = \text{estimate} \pm z^*\,SE$
Two-group SE	$\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}$

In the SE, p_i are sample proportions and q_i = 1 − p_i. Do not pool for an interval.

Check independence and at least ten observed successes and failures in each group.

Regression and residuals

Predict y from x	$\hat{y} = a + bx$
Least-squares slope	$b = r\,s_y/s_x$
Intercept	$a = \bar{y} - b\,\bar{x}$
Residual	$e = y - \hat{y}$

With an intercept, r^2 is the fraction of y variation explained by this line.

Use a residual plot to check the model and avoid extrapolating far beyond the data.

Statistics Formula Sheet

Conditional probability and Bayes

Given B	$P(A \mid B) = P(A \text{ and } B)/P(B)$
Reverse a conditional	$P(A \mid B) = P(B \mid A)P(A)/P(B)$
Split by a partition A_i	$P(B) = \sum P(B \mid A_i)P(A_i)$

Each conditioning event needs positive probability. Use all partition cases in the sum.

Counting and arrangements

Successive choices	**multiply available counts**
Order matters	$P(n,k) = n!/(n - k)!$
Order does not matter	$C(n,k) = n!/[k!(n - k)!]$

The formulas select k distinct objects from n, with 0 <= k <= n. Use 0! = 1.

Test statistics and decisions

Standardize	**(estimate − null)/SE**
One-proportion null SE	$\sqrt{(p_0(1 - p_0)/n)}$
One-mean t	$(\text{xbar} - \text{mu}_0)/(s/\sqrt{(n)})$
Reject H_0 if	**p-value ≤ significance level**

Use null SE for z and estimated SE for t.
For a two-proportion test of equality, pool counts.

Correlation and mean margins

Sample correlation	$r = \sum z_x z_y/(n - 1)$
Known population SD	$\text{xbar} \pm z^* \sigma/\sqrt{(n)}$
Margin of error	**critical value × standard error**

z_x and z_y are paired sample-standardized values.
For a mean interval, check independence and approximate normality of the mean.