Stat 101 Formulas

Sample Statistics

Sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$$

Sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

5-Number summary

$$Q_0 = minimum$$
$$Q_1 = 1st\ quartile$$
$$Q_2 = median$$
$$Q_3 = 3rd\ quartile$$
$$Q_4 = maximum$$

Range

$$Range = maximum - minimum$$
$$= Q_4 - Q_0$$

Inter-Quartile Range

$$IQR = Q_3 - Q_1$$

Fences for Outliers

$$Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR$$

Simple Linear Regression

Sample Covariance

$$Cov(x, y) = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)}$$

Sample Correlation

$$r = \frac{Cov(x, y)}{s_x s_y} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)s_x s_y}$$

Regression Model

$$\hat{y} = a + bx$$

Slope

$$b = r\frac{s_y}{s_x}$$

Intercept

$$a = \bar{y} - b\bar{x}$$

Residual

$$resid_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

Normal Distribution

Standardize

$$z = \frac{x - \mu}{\sigma}$$

Un-Standardize

$$x = \mu + z\sigma$$

68/95/99.7 Rule

$$P(-1 < Z < 1) \approx .68$$
$$P(-2 < Z < 2) \approx .95$$
$$P(-3 < Z < 3) \approx .997$$

kth Percentile

$$x\ such\ that\ P(X < x) = k\%$$

Stat 101 Formulas

Probability

Complement Rule

$$P(A^C) = 1 - P(A)$$

General Addition Rule

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

Multiplication Rule for Independent Events

$$P(A \cap B) = P(A) * P(B)$$

General Multiplication Rule

$$P(A \cap B) = P(A) * P(B|A) = P(B) * (P(A|B)$$

Conditional Probability

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

A and B are Independent if:

1) $P(A \cap B) = P(A) * P(B)$
2) $P(A) = P(A|B)$
3) $P(B) = P(B|A)$

Random Variables

Expected Value

$$\mu = E(X) = \sum_{i=1}^{k} x_i * P(X = x_i) = \sum_{i=1}^{k} x_i * p_i$$

Variance

$$\sigma^2 = Var(X) = E((X - \mu)^2) = E(X^2) - \mu^2 = \sum_{i=1}^{k} (x - \mu)^2 * p_i$$

Linearity of Expected Value

$$E(aX) = aE(X)$$
$$E(X + b) = E(X) + b$$
$$E(X + Y) = E(X) + E(Y)$$

Variance of a Linear Combination

$$Var(aX) = a^2 Var(X)$$
$$Var(X + b) = Var(X)$$
$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

Variance of Linear Combination of Independent X,Y

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$
$$SD(X_1 + X_2 + \cdots + X_n) = \sqrt{n} SD(X)$$

Stat 101 Formulas

Special Distributions
Bernoulli(p)

$$P(X = 1) = p$$
$$P(X = 0) = q = 1 - p$$
$$E(X) = p$$
$$Var(X) = pq$$

Binomial(n,p)
Sum of n independent Bernoullis

$$P(X = r) = \binom{n}{r} p^r q^{n-r} = \mathbf{binompdf(n, p, r)}$$
$$P(X \leq r) = \sum_{k=0}^{r} \binom{n}{k} p^k q^{n-k} = \mathbf{binomcdf(n, p, r)}$$

$$E(X) = np$$
$$Var(X) = npq$$

Central Limit Theorem
If $x_1, \ldots, x_n$ independent, come from a distribution with mean μ and standard deviation σ
$\bar{x}$ approximately follows a Normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Sampling Distributions (assuming CLT applies)
If $x_1, \ldots, x_n$ ~Bernoulli(p)

$$\sum x_i \sim Binom(n, p) \approx N(np, \sqrt{npq})$$
$$\hat{p} = \frac{\sum x_i}{n} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

If $x_1, \ldots, x_n$ ~ have mean μ and standard deviation σ

$$\sum x_i \sim N(n\mu, \sqrt{n}\sigma)$$
$$\bar{x} = \frac{\sum x_i}{n} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Confidence Intervals

(1-α)100% Confidence Interval
Estimate ± Margin of Error
Margin of Error = (# of Standard errors)*(Size of Standard Error)

Population proportion p (n large)	$\hat{p} \pm z_{\alpha/2} * \sqrt{\frac{\hat{p}\hat{q}}{n}}$
Population difference p_1-p_2 (n_1, n_2 large)	$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} * \sqrt{\frac{\hat{p}_1\hat{q}_1}{n} + \frac{\hat{p}_2\hat{q}_2}{n}}$
Population mean μ (n≥30, σ known)	$\bar{x} \pm z_{\alpha/2} * \frac{\sigma}{\sqrt{n}}$

Stat 101 Formulas

Population mean μ (n<30, σ known)	$\bar{x} \pm t_{\alpha/2} * \frac{\sigma}{\sqrt{n}}, \qquad t \sim T(n-1\, d.f.)$
Population mean μ (σ unknown)	$\bar{x} \pm t_{\alpha/2} * \frac{s}{\sqrt{n}}, \qquad t \sim T(n-1\, d.f.)$

Hypothesis Testing

$\alpha = P(Type\ I\ Error) = P(Reject\ H_0 | H_0\ is\ true)$

$\beta = P(Type\ II\ Error) = P(Don't\ reject\ H_0 | H_0\ is\ false)$

$Power = 1 - \beta$

Calculate the Test Statistic

Population proportion p (n large)	$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}}$
Population difference p_1-p_2 (n_1, n_2 large) H_0: p_1-p_2=0	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_{pooled}\hat{q}_{pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ And $\hat{p}_{pooled} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$
Population mean μ (n≥30, σ known)	$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
Population mean μ (n<30, σ known)	$t = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$
Population mean μ (σ unknown)	$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$

Conclusion

Test Type	p-value formula	calculator
Upper-Tail	$P(Z > z)$	normalcdf(z, 10)
Lower-Tail	$P(Z < z)$	normalcdf(−10, z)
Two-Tailed	$2 * P(Z > \|z\|)$	2*normalcdf(|z|, 10)

Reject H_0 if $p\text{-}value < \alpha$