INTRODUCTORY STATISTICS

Sheldon M. Ross

1 Introduction to Statistics

Statistics: the art of learning from data
Descriptive statistics: describes and summarizes data
Inferential statistics: draws conclusions from data
Population: collection of elements of interest
Sample: the part of the population from which data is obtained

2 Describing Data Sets

Frequency and relative frequency tables and graphs
Histograms
Stem-and-leaf plots
Scatter plots for paired data

3 Using Statistics to Summarize Data Sets

Sample mean: $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$

Sample median: the middle value

Sample variance: $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n - 1)$

Sample standard deviation: $s = \sqrt{s^2}$

Algebraic identity: $\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

Empirical rule for normal data sets:
- approximately 68% of the data lies within $\bar{x} \pm s$
- approximately 95% of the data lies within $\bar{x} \pm 2s$
- approximately 99.7% of the data lies within $\bar{x} \pm 3s$

Sample correlation coefficient:
$r = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})/[(n-1)s_x s_y]$

4 Probability

$0 \le P(A) \le 1$
$P(S) = 1$, where S is the set of all possible values
$P(A \cup B) = P(A) + P(B)$, when A and B are disjoint
Probability of the complement: $P(A^c) = 1 - P(A)$
Addition rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Conditional probability: $P(B|A) = P(A \cap B)/P(A)$
Multiplication rule: $P(A \cap B) = P(A)P(B|A)$
Independent events: $P(A \cap B) = P(A)P(B)$

5 Discrete Random Variables

Expected value (or mean): $E[X] = \sum_{i=1}^{n} x_i P\{X = x_i\}$

$E[X + Y] = E[X] + E[Y]$
Variance: $\mathrm{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$
Standard deviation: $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$
$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$ if X and Y are independent

Binomial random variable:

$$P\{X = i\} = \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i}, i = 0, \ldots, n$$

$E[X] = np \quad \mathrm{Var}(X) = np(1-p)$

6 Normal Random Variables

Normal random variable X: characterized by $\mu = E[X]$, $\sigma = \mathrm{SD}(X)$
Standard normal random variable Z: normal with $\mu = 0, \sigma = 1$
$P\{|Z| > x\} = 2P\{Z > x\}, x > 0$
$P\{Z < -x\} = P\{Z > x\}$
z_α is such that $P\{Z > z_\alpha\} = \alpha$
If X is normal then $Z = (X - \mu)/\sigma$ is standard normal.
Additive property: If X and Y are independent normals then $X + Y$ is normal with mean $\mu_x + \mu_y$, and variance $\sigma_x^2 + \sigma_y^2$

7 Distributions of Sampling Statistics

$X_1, \ldots, X_n$ is sample from population: $E[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2$
$E[\bar{X}] = \mu$
$\mathrm{Var}(\bar{X}) = \sigma^2/n$
Central limit theorem: $\sum_{i=1}^{n} X_i$ is, for large n, approximately normal with mean $n\mu$ and standard deviation $\sigma\sqrt{n}$; equivalently $\sqrt{n}(\bar{X} - \mu)/\sigma$ is approximately standard normal.

Normal approximation to binomial: If $np \ge 5, n(1-p) \ge 5$ then $[\mathrm{Bin}(n,p) - np]/\sqrt{np(1-p)}$ is approximately standard normal.

8 Estimation

$\bar{X}$ is the estimator of the population mean μ.
$\hat{p}$, the proportion of the sample that has a certain property, estimates p, the population proportion having this property.
S^2 estimates σ^2, and S estimates σ.
$100(1-\alpha)$ confidence interval estimator for μ:
- data normal or n large, σ known: $\bar{X} \pm z_{\alpha/2}\sigma/\sqrt{n}$
- data normal, σ unknown: $\bar{X} \pm t_{n-1,\alpha/2} S/\sqrt{n}$

$100(1-\alpha)$ confidence
interval for p: $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$

9 Testing Statistical Hypotheses

H_0 = null hypothesis: hypothesis that is to be tested

Significance level α: the (largest possible) probability of rejecting H_0 when it is true

p value: the smallest significance level at which H_0 would be rejected

Hypothesis Tests Concerning the Mean μ of a Population

Assumption: Either the distribution is normal or sample size n is large.

H_0	H_1	Test statistic TS	Significance-level-α test	p value if TS $= v$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\dfrac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$†	Reject H_0 if $\lvert TS \rvert \geq z_{\alpha/2}$	$2P\{z \geq \lvert v \rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\dfrac{\sqrt{n}(\overline{X} - \mu_0)}{\sigma}$†	Reject H_0 if $TS \geq z_\alpha$	$P\{Z \geq v\}$
$\mu = \mu_0$	$\mu \neq \mu_0$	$\dfrac{\sqrt{n}(\overline{X} - \mu_0)}{S}$	Reject H_0 if $\lvert TS \rvert \geq t_{n-1,\alpha/2}$	$2P\{T_{n-1} \geq \lvert v \rvert\}$
$\mu \leq \mu_0$	$\mu > \mu_0$	$\dfrac{\sqrt{n}(\overline{X} - \mu_0)}{S}$	Reject H_0 if $TS \geq t_{n-1,\alpha}$	$P\{T_{n-1} \geq v\}$

† *Assumption*: σ known.

Note: To test H_0: $\mu \geq \mu_0$, multiply data by -1 and use the above.

Hypothesis Tests Concerning p

(the proportion of a large population that has a certain characteristic)

X is the number of population members in a sample of size n that have the characteristic. B is a binomial random variable with parameters n and p_0.

H_0	H_1	Test statistic TS	p value if TS $= x$
$p \leq p_0$	$p > p_0$	X	$P\{B \geq x\}$
$p = p_0$	$p \neq p_0$	X	$2 \text{ Min } \{P\{B \leq x\}, P\{B \geq x\}\}$

10 Hypotheses Tests Concerning Two Populations

Tests Concerning the Means of Two Populations When Samples Are Independent

The X sample of size n and the Y sample of size m are independent.

H_0	H_1	Test statistic TS	Assumptions	Significance level α test	p value if TS $= v$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	n, m large	Reject if $\lvert TS \rvert \geq z_{\alpha/2}$	$2P\{Z \geq \lvert v \rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{S_x^2/n + S_y^2/m}}$	n, m large	Reject if $TS \geq z_\alpha$	$P\{Z \geq v\}$
$\mu_x = \mu_y$	$\mu_x \neq \mu_y$	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Normal populations $\sigma_x = \sigma_y$	Reject if $TS \geq t_{n+m-2,\alpha/2}$	$2P\{T_{n+m-2} \geq \lvert v \rvert\}$
$\mu_x \leq \mu_y$	$\mu_x > \mu_y$	$\dfrac{\overline{X} - \overline{Y}}{\sqrt{S_p^2(1/n + 1/m)}}$	Normal populations $\sigma_x = \sigma_y$	Reject if $TS \geq t_{n+m-2,\alpha}$	$P\{T_{n+m-2} \geq v\}$

$$S_p^2 = \frac{n-1}{n+m-2} S_x^2 + \frac{m-1}{n+m-2} S_y^2 = \text{pooled estimator of } \sigma_x^2 = \sigma_y^2$$

Tests Concerning Two Population Proportions

p_1 and p_2 are the proportions of the members of two populations that have a certain characteristic. A random sample of size n_1 is chosen from the first population, and an independent random sample of size n_2 is chosen from the second. $\hat{p}_1$ and $\hat{p}_2$ are the proportions of the samples that have the characteristic and $\hat{p}$ is the proportion of the combined samples that has it.

H_0	H_1	Test statistic TS	Significance-level-α test	p value if TS $= v$
$p_1 = p_2$	$p_1 \neq p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1-\hat{p})}}$	Reject H_0 if $\lvert TS\rvert \geq z_{\alpha/2}$	$2P\{Z \geq \lvert v\rvert\}$
$p_1 \leq p_2$	$p_1 > p_2$	$\dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{(1/n_1 + 1/n_2)\hat{p}(1-\hat{p})}}$	Reject H_0 if TS $\geq z$	$P\{Z \geq v\}$

11 Analysis of Variance

One-Factor ANOVA Table

$\overline{X}_i$ and S_i^2, $i = 1, \ldots, m$, are the sample means and sample variances of independent samples of size n from normal populations having means μ_i and a common variance σ^2.

Source of estimator	Estimator of σ^2	Value of test statistic
Between samples	$n\overline{S}^2 = \dfrac{n\sum_{i=1}^{m}(\overline{X}_i - \overline{\overline{X}})^2}{(m-1)}$	$\text{TS} = \dfrac{n\overline{S}^2}{\left(\sum_{i=1}^{m} S_i^2\right)\Big/ m}$
Within samples	$\left(\sum_{i=1}^{m} S_i^2\right)\Big/ m$	

Significance-level-α test of H_0: all μ_i are equal

Reject H_0 if TS $\geq F_{m-1,\, m(n-1),\alpha}$

Do not reject otherwise

If TS $= v$ then

$$p \text{ value} = P\{F_{m-1,m(n-1)} \geq v\}$$

where $F_{m-1,m(n-1)}$ is an F random variable with $m-1$ numerator and $m(n-1)$ denominator degrees of freedom.

Two-factor ANOVA model: For $i = 1, \ldots, m,\ j = 1, \ldots, n$

$$E[X_{ij}] = \mu + \alpha_i + \beta_j$$

$$\sum_{i=1}^{m}\alpha_i = \sum_{j=1}^{n}\beta_j = 0$$

μ is the grand mean, α_i is the deviation from the grand mean due to row i, and β_j is the deviation from the grand mean due to column j. Their estimators are

$$\mu = X.. \quad \hat{\alpha}_i = X_i. - X.. \quad \hat{\beta}_j = X._j - X..$$

Two-Factor ANOVA Table

	Sum of squares	Degrees of freedom
Row	$SS_r = n\sum_{i=1}^{m}(X_i. - X..)^2$	$m-1$
Column	$SS_c = m\sum_{j=1}^{n}(X._j - X..)^2$	$n-1$
Error	$SS_e = \sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij} - X_i. - X._j + X..)^2$	$N = (n-1)(m-1)$

Null hypothesis	Test statistic	Significance-level-α test	p value if TS $= v$
No row effect (all $\alpha_i = 0$)	$\dfrac{SS_r/(m-1)}{SS_e/N}$	Reject if TS $\geq F_{m-1,N,\alpha}$	$P\{F_{m-1,N} \geq v\}$
No column effect (all $\beta_j = 0$)	$\dfrac{SS_c/(n-1)}{SS_e/N}$	Reject if TS $\geq F_{n-1,N,\alpha}$	$P\{F_{n-1,N} \geq v\}$

12 Linear Regression

Simple linear regression model: $Y = \alpha + \beta x + e$

Least square estimators: $\hat{\beta} = S_{xY}/S_{xx}$, $\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$

$$S_{xY} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{1}^{n}x_iY_i - n\overline{x}\overline{Y}$$

$$S_{xx} = \sum_{1}^{n}(x_i - \overline{x})^2 = \sum_{1}^{n}x_i^2 - n\overline{x}^2$$

Estimated regression line: $y = \hat{\alpha} + \hat{\beta}x$

Error term e is normal with mean 0 and variance σ^2. Estimator of σ^2 is $SS_R/(n-2)$, $SS_R = \sum_i (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 = (S_{xx}S_{YY} - S_{xY}^2)/S_{xx}$

To test $H_0\colon \beta = 0$. Use $TS = \sqrt{(n-2)S_{xx}/SS_R}\,\hat{\beta}$

Significance-level-γ test is to reject H_0 if $|TS| \geq t_{n-2,\gamma/2}$.
If $TS = v$, p value $= 2P\{T_{n-2} \geq v\}$
$100(1-\gamma)$ confidence prediction interval for response at input x_0

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2,\gamma/2}\sqrt{(1 + 1/n + (x_0 - \bar{x})^2/S_{xx})SS_R/(n-2)}$$

Coefficient of determination: $R^2 = 1 - SS_R/S_{YY}$ is the proportion of the variation in the response variables that is explained by the different input values. Its square root is the absolute value of the sample correlation coefficient.

Multiple linear regression model:
$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$

13 Chi-squared Goodness-of-Fit Tests

P_i is the proportion of population with value i, $i = 1, \ldots, k$.

To test $H_0\colon P_i = p_i$, $i = 1, \ldots, k$, take a sample of size n. Let N_i be the number equal to i, $e_i = np_i$, $TS = \sum_{i=1}^{k}(N_i - e_i)^2/e_i$. Significance-level-$\alpha$ test rejects H_0 if $TS \geq \chi^2_{k-1,\alpha}$.

If $TS = v$, then p value $= P\left\{\chi^2_{k-1} \geq v\right\}$.

Suppose each member of a population has an X and a Y characteristic. Assume r possible X and s possible Y characteristics. To test for independence of the characteristics of a randomly chosen member, choose a sample of size n.

N_{ij} = number with X characteristic i and Y characteristic j
N_i = number with X characteristic i
M_j = number with Y characteristic j $\qquad \hat{e}_{ij} = N_i M_j/n$

If $\sum_i\sum_j(N_{ij} - \hat{e}_{ij})^2/\hat{e}_{ij} \geq \chi^2_{(r-1)(s-1),\alpha}$ then the hypothesis of independence is rejected at significance level α.

14 Nonparametric Hypotheses

Let η = median of population. The *sign* test of

$H_0\colon \eta = m$ against $H_1\colon \eta \neq m$

takes a sample of size n. If i are less than m, then

p value $= 2\,\text{Min}\,(P\{N \leq i\}, P\{N \geq i\})$

where N is a binomial $(n, 1/2)$ random variable.

The *signed rank* test is used to test the hypothesis that a population distribution is symmetric about 0. It ranks the data in terms of absolute value. TS is the sum of the ranks of the negative values. If $TS = t$, then

p value $= 2\,\text{Min}\,(P\{TS \leq t\}, P\{TS \geq t\})$

TS is approximately normal with mean $n(n+1)/4$ and variance $n(n+1)(2n+1)/24$.

To test equality of two population distributions, draw random samples of sizes n and m and rank the $n + m$ data values. The *rank sum* test uses TS = sum of ranks of first sample. It rejects H_0 if TS is either significantly large or significantly small. If $TS = t$, then

p value $= 2\,\text{Min}\,(P\{TS \leq t\}, P\{TS \geq t\})$

TS is approximately normal with mean $n(n+m+1)/2$ and variance $nm(n+m+1)/12$.

To test the hypothesis that a sequence of 0s and 1s is random, use the *runs* test by counting R, the number of runs. Reject randomness when R is either too small or too large to be explained by chance. Use the result that when H_0 is true, R is approximately normal with mean $1 + 2\,nm/(n+m)$ and variance

$$\frac{2\,nm(2\,nm - n - m)}{(n+m)^2(n+m-1)}$$

15 Quality Control

Control chart limits $\mu \pm 3\sigma/\sqrt{n}$ $\qquad$ n = subgroup size

Area under the Standard Normal Curve to the Left of *x*

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5733
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998