Formula Sheet for Final Exam

STAT 011

Sample Statistics

For a sample of data

If $\{x_1, x_2, \ldots, x_n\}$ is a data set of n observational units, we have the following:

Sample mean

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

Sample variance

$$Var(x_1, \ldots, x_n) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sample standard deviation

$$sd(x_1, \ldots, x_n) = s = \sqrt{s^2}$$

If we want to standardize the data set X, to create a new standardized data set $Z = \{z_1, z_2, \ldots, z_n\}$ we preform

$$z_i = \frac{x_i - \bar{x}}{sd(x_1, \ldots, x_n)}, \text{ for } i = 1, \ldots, n.$$

Simple linear regression notation

The fitted/estimated regression model is $\hat{y}_i = b_0 + b_1 x_i$ where $b_0 = \bar{y} - b_1\bar{x}$ and $b_1 = \frac{s_{xy}}{\sqrt{s_x s_y}} \cdot \frac{s_y}{s_x}$.

$$\text{Residual} = e = y - \hat{y} = \text{observed value} - \text{predicted value}$$

Standard error of the residuals: $s_e = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$

Sum of squares terms

$$s_x = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad s_y = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Correlation coefficient

$$r = \frac{s_{xy}}{\sqrt{s_x s_y}}$$

Probability

Five Laws of Probability

1) A probability is a number between 0 and 1.

$$0 \leq Pr(A) \leq 1, \quad \text{for } A \in S$$

2) The probability of the set of all possible outcomes of a trial is 1.

$$Pr(S) = 1$$

3) The probability of an event not occuring is equal to 1 minus the probability the event does occur.

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

4) For any events in the sample space of a random variable, say, A and B, we compute the probability of event A or event B or both events A and B occurring with the formula:

$$Pr(A \text{ or } B) = Pr(A) + Pr(B) - Pr(A \text{ and } B)$$

5) If an event A is independent of another event B, then the probability that both events occur is the product of the probabilities of the two individual events:

$$Pr(A \text{ and } B) = Pr(A) \times Pr(B).$$

Definition of conditional probability

$$Pr(B \mid A) = \frac{Pr(A \text{ and } B)}{Pr(A)}$$

General multiplication rule

For any random events A and B (that need not be independent),

$$Pr(A \text{ and } B) = Pr(A) \times Pr(B \mid A).$$

Law of total probability

$$Pr(B) = Pr(B \text{ and } A) + Pr(B \text{ and } A^C)$$

Random Variables

For a random variable X,

$$E(X) = \sum_{x \in S} \left[x \times Pr(x)\right], \quad Var(X) = \sum_{x \in S} \left[(x - E(X))^2 \times Pr(x)\right], \quad st.dev(X) = \sqrt{Var(X)}.$$

For two random variables, X and Y:

$$Cov(X,Y) = E\left[(X - E(X)) \cdot (Y - E(Y))\right], \quad Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}.$$

Linear transformations of a random Variable

Suppose a is some number between $-\infty$ and $+\infty$. The following are properties of expectation and variance for linear transformations of a random variable X.

- $E(aX) = aE(X), \quad E(a \pm X) = a \pm E(X)$
- $Var(aX) = a^2Var(X), \quad Var(a \pm X) = Var(X)$

Linear transformations of two random variables

Suppose both X and Y are random variables that may or may not be related to one another. The following are properties of expectation and variance for linear transformations involving both random variables.

- $E(X \pm Y) = E(X) \pm E(Y)$
- $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$
- If X and Y are independent random variables, then $Cov(X,Y) = 0$.

Normal Random Variable

If $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

Binomial Random Variable

If $X \sim Bin(n,p)$ then $Pr(X = x) = nCx \cdot p^x \cdot (1-p)^{n-x}$, where $nCx = \frac{n!}{x!(n-x)!}$.

Sampling Distributions

Under appropriate conditions, the sampling distribution for the sample proportion is

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

The standard error for the sample proportion is $SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Under appropriate conditions, the sampling distribution for the sample mean is

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

The standard error for the sample mean is $SE(\bar{x}) = \frac{s}{\sqrt{n}}$.

Confidence Intervals

For a single proportion

$$\hat{p} \pm [z_a^* \times SE(\hat{p})]$$

where z_a^* is the lower (or upper) $\left(\frac{1-a}{2}\right)^{th}$ quantile of a $N(0,1)$ distribution for confidence level a.

For a single mean

$$\bar{x} \pm [t_{a,(n-1)}^* \times SE(\bar{x})]$$

where $t_{a,(n-1)}^*$ is the lower (or upper) $\left(\frac{1-a}{2}\right)^{th}$ quantile of a t-distribution with $n-1$ degrees of freedom, for confidence level a.

For a difference in proportions

$$(\hat{p}_1 - \hat{p}_2) \pm [z_a^* \times SE(\hat{p}_1 - \hat{p}_2)]$$

where $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ and z_a^* is the lower (or upper) $\left(\frac{1-a}{2}\right)^{th}$ quantile of a $N(0,1)$ distribution for confidence level a.

For a difference in means

Independent samples

$$(\bar{x}_1 - \bar{x}_2) \pm [t_{a,(\nu)}^* \times SE(\bar{x}_1 - \bar{x}_2)]$$

where $SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and $t_{a,(\nu)}^*$ is the lower (or upper) $\left(\frac{1-a}{2}\right)^{th}$ quantile of a t-distribution with ν degrees of freedom, for confidence level a. (These degrees of freedom will always be provided to you as they are complicated to derive.)

Paired samples

$$\bar{d} \pm [t_{a,(n-1)}^* \times SE(\bar{d})]$$

where $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$ and $t_{a,(n-1)}^*$ is the lower (or upper) $\left(\frac{1-a}{2}\right)^{th}$ quantile of a t-distribution with $n-1$ degrees of freedom, for confidence level a.

Hypothesis Tests

For a single proportion

We can test $H_0 : p = p_0$ with the test statistic $T.S. = \frac{\hat{p}-p_0}{st.dev(\hat{p})}$, where $st.dev(\hat{p}) = \sqrt{\frac{p_0(1-p_0)}{n}}$.

For a single mean

We can test $H_0 : \mu = \mu_0$ with the test statistic $T.S. = \frac{\bar{x}-\mu_0}{SE(\bar{x})}$.

For a difference in proportions

We can test $H_0 : p_1 - p_2 = 0$ with the test statistic $T.S. = \frac{(\hat{p}_1-\hat{p}_2)-0}{SE(\hat{p}_1-\hat{p}_2)}$.

For a difference in means

Independent samples

We can test $H_0 : \mu_1 - \mu_2 = \Delta_0$ with the test statistic $T.S. = \frac{(\bar{x}_1-\bar{x}_2)-\Delta_0}{SE(\bar{x}_1-\bar{x}_2)}$.

Paired samples

We can test $H_0 : \mu_d = \Delta_0$ with the test statistic $T.S. = \frac{\bar{d}-\Delta_0}{SE(\bar{d})}$.

For count data

The chi-square goodness of fit test tests the null $H_0 : p_1 = p_{1,0}, p_2 = p_{2,0}, p_3 = p_{3,0}, \ldots, p_k = p_{k,0}$ with the test statistic $T.S. = \sum_{\text{all cells}} \frac{(Obs-Exp)^2}{Exp}$.

The chi-square test of homogeneity tests the null $H_0 : p_1 = p_2 = p_3 = \cdots = p_k$ with the test statistic $T.S. = \sum_{\text{all cells}} \frac{(Obs-Exp)^2}{Exp}$.

The chi-square test of independence tests the null H_0 : Variable X is independent of variable Y with the test statistic $T.S. = \sum_{\text{all cells}} \frac{(Obs-Exp)^2}{Exp}$.