Statistics Formulas

Prob-Stats, Math 3550 (Sinn)

1 Producing Data

1.1 Probability

Probability rules:

- For any event A, $0 \leq P(A) \leq 1$
- The sample space S has probability $P(S) = 1$
- For disjoint event sets A and B,

$$P(A \text{ or } B) = P(A) + P(B)$$

- In general, for event sets A and B,

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

- $P(A \text{ does not occur}) = 1 - P(A)$
- For a discrete probability density function $p(x)$:
  - $0 \leq p(x_i) \leq 1 \quad \text{for all} \quad 1 \leq i \leq n$
  - $\sum p(x_i) = 1$
- For any continuous probability density function $f(x)$:
  - $f(x) \geq 0 \quad \text{for all} \quad x \in (-\infty, +\infty)$
  - $\int\limits_{-\infty}^{+\infty} f(x)dx = 1$
- The expected value of the probability density function:

$$E(X) = \sum p(x_i)x_i = \mu_x$$

or

$$E(X) = \int\limits_{-\infty}^{+\infty} xf(x)dx = \mu_x$$

- The variance of a pdf is defined as $E[(X - \mu_x)^2]$ but is more easily computed as:

$$E(X) = E(X^2) - E(X)^2$$

- The standard deviation squared is equal the variance: $\sigma^2 = V(X)$

---

2 Exploring Data: Distributions and Descriptives

Look for overall pattern (shape, center, spread) and deviations (outliers).

- Mean:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n}\sum x_i$$

- Standard deviation:

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

- Median: Arrange all observations from smallest to largest. The median M is located $\frac{(n+1)}{2}$ from the beginning of this list. We also use the notation $\tilde{x}$ for the median.

- Quartiles: The first quartile $Q1$ is the median of the observations whose position in the ordered list is to the left of the location of the overall median. The third quartile $Q3$ is the median of the observations to the right of the location of the overall median.

- Five-number summary:

Minimum	Min
1st Quartile	Q1
Median	$\tilde{x}$
3rd Quartile	Q3
Maximum	Max

- Standardized value of x:

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

3 Exploring Data: Relationships

Look for overall pattern (form, direction, strength) and deviations (outliers, influential observations).

- Correlation (conceptual form using z-scores):

$$\begin{aligned} r =& \frac{1}{n-1} \sum z_{x_i} z_{y_i} \\ =& \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right) \end{aligned}$$

- Least-squares regression line:

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

$$\begin{array}{ll} \text{with slope} & b = r \cdot \frac{s_y}{s_x} \\ \text{and intercept} & a = \bar{y} - b\bar{x} \\ \text{thus containing} & (\bar{x}, \bar{y}) \end{array}$$

  The slope b of the regression line (or line of best fit) is the standard change (deviation) in y over the standard change (deviation) in x multiplied by the strength (correlation) of the relationship between the x- and y-variables.

- Residuals: for any data point (x_i, y_i):

$$\text{Residual} = \text{Observed } y_i - \text{Predicted } y_i$$

or

$$y_i - \hat{y}_i = y_i - a - bx_i$$

which gives the vertical distance between the actual y-value and the line of best fit. The "line of best" is the "least squares line" that minimizes the sum of the squared residuals, e.g. minimizes the total vertical distance between the actual data and the line of best fit (using calculus).

4 Sampling distribution of a sample mean:

- $\bar{x}$ has mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$
- $\bar{x}$ has a Normal distribution if the population distribution is Normal.
- Central limit theorem: X is approximately Normal when n is large.

5 Basics of Inference

- z confidence interval for a population mean (σ known, SRS from Normal population):

$$\mu \in \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \qquad \text{with } z^* \text{ from } N(0,1)$$

- Sample size for desired margin of error m:

$$n = \left(\frac{z^* \sigma}{m} \right)^2$$

- z test statistic for $H_0 : \mu = \mu_0$ (σ known, SRS from Normal population, rarely used in modern practice):

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

with P-values from $N(0, 1)$.

5.1 Inference About Means

- The t confidence interval for a population mean (SRS from Normal Population):

$$\mu \in \bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

with t^* from the t-distribution with degrees of freedom $n - 1$.

- t test statistic $H_0 : \mu = \mu_0$ (SRS from Normal Population):

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

with P-values from the t-distribution with degrees of freedom$n - 1$.

- Matched pairs: To compare the responses to the two treatments, apply the one-sample t procedures to the observed differences

- Two-sample confidence interval for $\mu_1 - \mu_2$ (independent SRSs from the Normal Populations):

$$(\mu_1 - \mu_2) \in (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

with conservative t^* from t with $df = \min(n_1 - 1, n_2 - 1)$, or use software.

- Two-sample t test statistic for $H_0 : \mu_1 = \mu_2$ (independent SRSs from Normal populations):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

with conservative P-values from t with $df = \min(n_1 - 1, n_2 - 1)$, or use software.

5.2 Inference About Proportions

- Sampling distribution of a sample proportion: when the population and the sample size are both large and p is not close to 0 or 1, $\hat{p}$ is approximately $N\left(p, \sqrt{p(1-p)/n}\right)$.

- Large-sample z confidence interval for p:

$$\text{prop} \in \hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

with z^* from $N(0, 1)$. Plus Four Method greatly improves accuracy: same formula after adding four imaginary observations: two success and two failures.

- The z test statistic for $H_0 : p = p_0$ (large SRS):

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \qquad \text{with } P\text{-values from } N(0,1)$$

- Sample size for desired margin of error m:

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$$

where p^* is a guessed value for p, or $p^* = 0.5$

- Large-sample z confidence interval for $p_1 - p_2$:

$$(p_1 - p_2) \in (\hat{p}_1 - \hat{p}_2) \pm z^* \text{SE}$$

with z^* from $N(0, 1)$ and standard error:

$$\text{SE} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Plus four to greatly improve accuracy: use the same formulas after adding one success and one failure to each sample.

- Two-sample z test statistic for $H_0 : p_1 = p_2$ (large independent SRSs):

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where $\hat{p}$ is the pooled (overall) proportion of successes.

6 The χ^2 Test

- Calculating cells of Expected Matrix:

$$\text{expected count} = \frac{\text{row total } \times \text{ column total}}{\text{table total}}$$

- χ^2 test statistic for testing whether the row and column variables in an $r \times c$ table are unrelated (expected cell counts not too small):

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

with P-values from the χ^2 distribution with

$$df = (r-1) * (c-1)$$

7 Inference for Regression

- Conditions for regression inference: n observations on x and y. The response y for any fixed x has a Normal distribution with mean given by the true regression line $y = \alpha + \beta x$ and standard deviation σ. Parameters are α, β, σ.
- Estimate α by the intercept a and β by the slope b of the least-squares line. Estimate σ by the regression standard error:

$$s = \sqrt{\frac{1}{n-2} \sum \text{residual}^2}$$

- A t confidence interval for regression slope β can be calculated by hand as follows:

$$b \pm t^* \text{SE}_b \qquad t^* \text{ from } t(n-2)$$

where

$$SE_b = \frac{\sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

However, best practice strongly indicates using software for all standard errors in regression.

- Testing for no correlation, $H_0 : \rho = 0$:

$$t = r \cdot \sqrt{\frac{n-2}{1-r^2}} \qquad t^* \text{ from } t(n-2)$$

where ρ is the parameter that r estimates.

8 One-way Analysis of Variance: Comparing Several Means

- ANOVA tests whether k populations have the same mean based on independent SRS's from k normal populations. The p-values come from the F distribution with $k-1$ and $N-k$ degrees of freedom, where N is the total observations in all samples.
- Describe the data using the k sample means $(\bar{x}_i)$ and standard deviations (s_i) and side-by-side graphs of the samples. The overall sample size is $N = n_1 + n_2 + \ldots + n_k$, and the grand mean is $(\bar{x})$, the airthmetic average of all N observations.
- The F test statistic is given by

$$F = \frac{\text{MSB}}{\text{MSW}} \qquad \text{or} \qquad F = \frac{\text{MSG}}{\text{MSE}}$$

where MSB is "between group" mean sum of squares (or MS Factor for TI calculators):

$$\text{MSB} = \frac{n_1(\bar{x}_1 - \bar{x})^2 + ... + n_k(\bar{x}_k - \bar{x})^2}{k-1}$$

and MSW is "within group" mean sum of squares and (or MSE for mean sum of squares for the "error" on TI):

$$\text{MSW} = \text{MSE} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{i_j} - \bar{x}_i)^2}{N-k}$$

However, with some algebra, this reduces to a more computationally friendly version:

$$\text{MSW} = \frac{(n_1 - 1)s_1^2 + ... + (n_k - 1)s_k^2}{N-k}$$

8.1 *Post Hoc* Testing

After finding significant differences between the sample means, we employ a *post hoc* test to ferret out significantly different group means. Tukey's HSD (Honestly Significant Difference) is the most common. However, Tukey's HSD is liberal and tends to err on the side of finding significance differences. Using Scheffe's or Dunnet's may be preferable where a more conservative or flexible approach is appropriate.

- Tukey's HSD must be computed for each individual pair of group means and depends upon the harmonic mean (see below). The HSD value is given by:

  $$\text{HSD}_{ij} = q^* \sqrt{\frac{\text{MSW}}{n_{ij}}}$$

  where q^* is a value from the Studentized Range Statistic (found in a table based upon α and degrees of freedom).

- We flag as significantly different any pair of groups i and j such that

  $$|\bar{x}_i - \bar{x}_j| > HSD_{ij}$$

8.2 Harmonic Mean

The harmonic mean H is the reciprocal of the arithmetic mean of the reciprocals. For 2 real numbers r and s we have

$$H_{rs} = \left(\frac{\frac{1}{r} + \frac{1}{s}}{2}\right)^{-1}$$

which simplifies to

$$H_{rs} = \frac{2rs}{r+s}$$

In the HSD calculation, n_{ij} refers to the harmonic mean of group sizes n_i and n_j:

$$n_{ij} = \frac{2n_i n_j}{n_i + n_j}$$