DATA1220-55 Cheat Sheet

Sarah E. Grabinski

2024-11-11

Formulas

Sample Proportion

$$\hat{p} = \frac{\text{count (something)}}{\text{count (everything)}} = \frac{\text{count (something)}}{n}$$

Sample Mean

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

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Sample Variance

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

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Sample Standard Deviation

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

Table 1: Standard Errors of Sample Statistics

Measure	SE	Calculation
Mean	$\text{SE}_{\bar{x}}$	$\frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$
Paired Difference in Means	$\text{SE}_{\bar{x}_{\text{difference}}}$	$\frac{\sigma_{\text{difference}}}{n} \approx \frac{s_{\text{difference}}}{n}$
Difference in Means	$\text{SE}_{\bar{x}_1 - \bar{x}_2}$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Proportion	$\text{SE}_{\hat{p}}$	$\sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
Difference in Proportions ($H_0: p_1 - p_2 = \mu$)	$\text{SE}_{\hat{p}_1 - \hat{p}_2}$	$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \approx \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$
Difference in Proportions ($H_0: p_1 - p_2 = 0$)	$\text{SE}_{\hat{p}_1 - \hat{p}_2}$	$\sqrt{\frac{p_{\text{pool}}(1-p_{\text{pool}})}{n_1} + \frac{p_{\text{pool}}(1-p_{\text{pool}})}{n_2}} \approx \sqrt{\frac{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}})}{n_1} + \frac{\hat{p}_{\text{pool}}(1-\hat{p}_{\text{pool}})}{n_2}}$

Terms

Population The entire group being researched (e.g. sample, study, target)

Sample A subset of the population, ideally random and large enough to be representative

Sample size The total number of subjects or observations in the sample, represented by n.

Reliability The consistency of the observed measurements from a sample. Data from a sample is considered reliable estimate of the sample statistic when there is very little bias or measurement error.

Validity The degree to which the sample statistic approximates the population parameter. A sample statistic is considered a valid estimate of the population parameter when the sample is large and/or representative of the study population.

Median The middle value in the data separating the top 50% from the bottom 50%. Found by arranging all values from lowest to highest and taking the middle value (or mean of the 2 middle values)

Quartile Each of the 4 equal groups into a which a population can be divided. The divisions between the quartiles are Q1 = 0.25 (25th percentile), Q2 = 0.50 (50th percentile, median), and Q3 = 0.75 (75th percentile).

Interquartile Range (IQR) The difference between the 3rd quartile (Q3 = 0.75) and the 1st quartile (Q1 = 0.25). The middle 50% of the data.

Mean Also called the average. The sum of all values in the sample divided by number of values in the sample. μ (mu) represents the mean of a population, and $\bar{x}$ represents the mean of a sample.

Variance Dispersion (spread) around the mean, determined by averaging the squared differences of all values from the mean. σ^2 (sigma squared) represents the variance of a population, and s^2 represents the variance of a sample.

Standard Deviation The square root of the variance. Also measures dispersion (spread) around the mean, but in the same units as the variable. σ (sigma) represents the standard deviation of a population, and s represents the standard deviation of a sample.

Central Limit Theorem The distribution of a sample statistic approximates the normal distribution $N\,(\text{population parameter}, \text{standard error})$ as $n \to \infty$.

Sampling Distribution the distribution of theoretically possible sample statistics from all samples of size n that can be taken from a population

Standard Error The standard deviation of a sampling distribution. Reflects how variable a sample statistic is expected to be from sample to sample.

Confidence Interval A range of values within which you expect the "true" population parameter to fall if you repeated the study an infinite number of times. The confidence level is the percentage of samples whose confidence interval would capture the "true" population parameter. A confidence intervals upper and lower bounds are found by calculating point estimate $\pm$ critical value $\times$ standard error.

Critical Value The number which defines the upper and lower bounds of a confidence interval from a given distribution. Its value corresponds to the probabilities $\alpha/2$ and $1 - \alpha/2$.

Null Hypothesis There is no meaningful relationship in the data. Represented as H_0, gives the null distribution under which the hypothesis is tested.

Alternate Hypothesis There is something meaningful in the data. Represented as H_A, indicates whether the hypothesis test is one-sided (left- or right-tailed) or two-sided (both tails).

Test Statistic The standardized value of the observed sample statistic under the null hypothesis H_0, used to find the p-value of a hypothesis test.

Type I Error The probability of rejecting the null hypothesis H_0 when H_0 is actually "true." Represented by α.

Type II Error The probability of failing to reject the null hypothesis H_0 when H_0 is not actually "true."

Inference

Means

Table 2: Sample Statistics for Inference of Population Means

Measure	Sample Statistic	Population Parameter
Mean	$\bar{x}$	μ
Paired Difference in Means	$\bar{x}_{\text{difference}}$	$\mu_{\text{difference}}$
Difference in Means	$\bar{x}_1 - \bar{x}_2$	$\mu_1 - \mu_2$
Standard Deviation	s	σ

Assumptions

- ***Independence***: sample observations are independent (i.e. random sample).
- ***Sample size***: the sample size should be greater than 30 ($n \geq 30$) with no extreme outliers.
- ***Normality***: when the sample size n is small, observations come from a normally distribution population. This condition relaxes as $n \to \infty$.
- ***Validity***: sample statistics approximate the population parameters ($\bar{x} \approx \mu$, $s \approx \sigma$)

Single Mean ($\bar{x}$) - One-Sample t-test

- A ***1-sample*** t-test tests if the mean (μ) of a population is different from a null value (μ_0).
- Sample statistics $\bar{x}$ (mean) and s (standard deviation) and the sample size n are used to infer the sampling distribution of the mean $\bar{x} \sim \mathrm{N}\left(\mu, SE_{\bar{x}}\right)$.
- To account for using $s \approx \sigma$ in the standard error, confidence intervals and hypothesis tests are based on the T distribution (Student's t) with the parameter degrees of freedom (df) $= n - 1$.

Confidence Interval

The confidence interval for the mean $\bar{x}$ estimating μ is…

$$\bar{x} \pm T^*_{\text{df}} \times SE_{\bar{x}}$$

The standard error of $\bar{x}$ is estimated by…

$$\text{SE}_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$$

The critical value from the t distribution with degrees of freedom $\text{df} = n - 1$ is…

$$T^*_{\text{df}} = T_{\text{df},\alpha/2} = T_{\text{df},1-\alpha/2}$$

A critical value is calculated from the t distribution in R using the function `qt()`. This function takes a probability `p` ($\alpha/2$ or $1 - \alpha/2$) and degrees of freedom `df` ($n - 1$).

```
qt(alpha/2, df=n-1)

qt(1-alpha/2, df=n-1)
```

Hypothesis Test

The null hypothesis of a 1-sample t-test states that the population mean μ is equal to some null value μ_0.

$$H_0 : \mu = \mu_0$$

The null distribution of $\bar{x}_0$ given the null hypothesis that $\mu = \mu_0$ is $\bar{x}_0 \sim \text{N}\left(\mu_0, SE_{\bar{x}_0}\right)$. The standard error of $\bar{x}_0$ is estimated as…

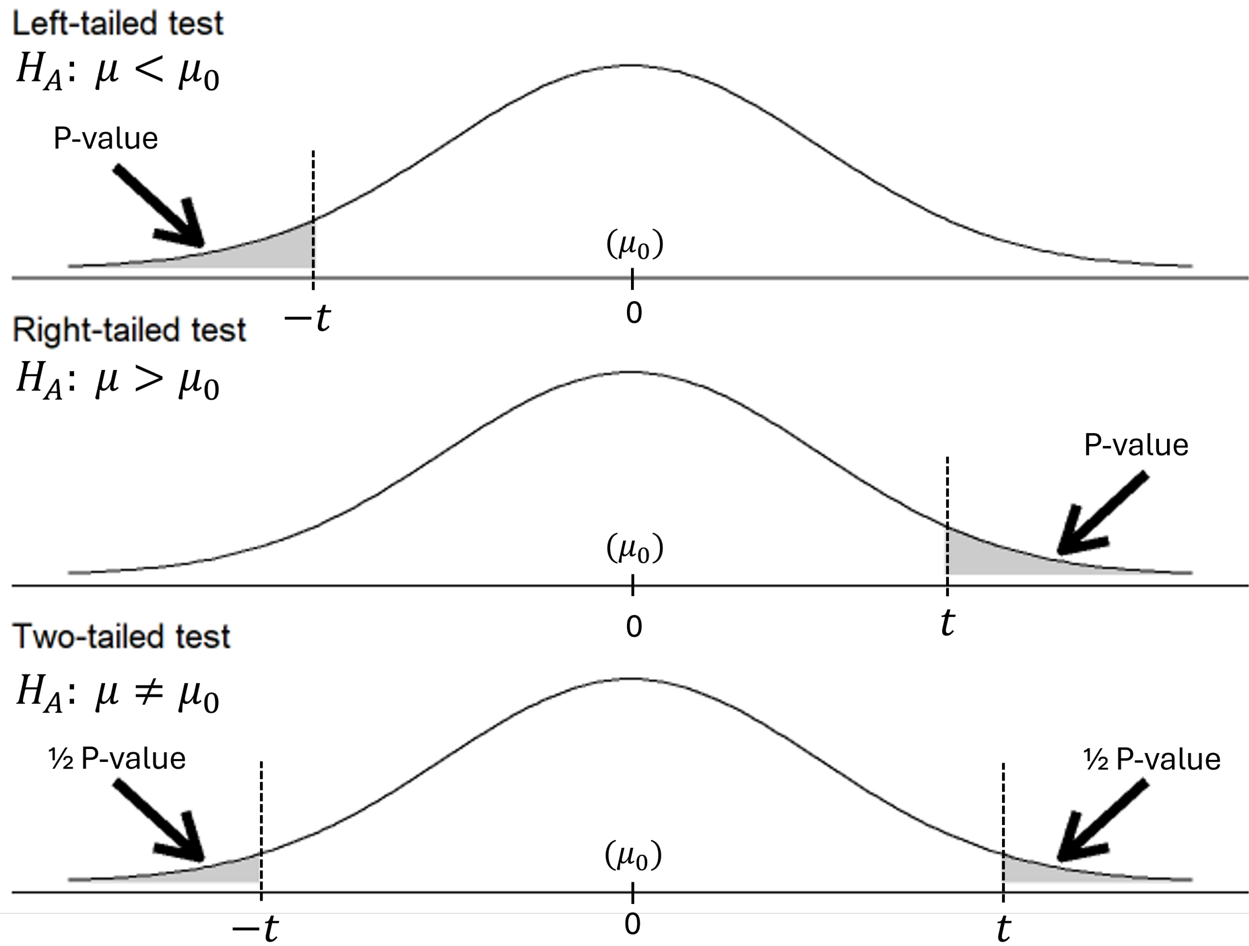
$$\text{SE}_{\bar{x}_0} = \frac{s}{\sqrt{n}}$$

A confidence interval for the mean under the null hypothesis is calculated as above, but using μ_0 as the point estimate.

$$\mu_0 \pm T^*_{\text{df}} \times \text{SE}_{\bar{x}_0}$$

The alternate hypothesis of a 1-sample t-test states that the population mean μ is greater than, less than, or not equal to some null value μ_0.

- $H_A : \mu < \mu_0$, left-tailed test (one-sided)
- $H_A : \mu > \mu_0$, right-tailed test (one-sided)
- $H_A : \mu \neq \mu_0$, two-tailed test (two-sided)



The test statistic t (the t-statistic) calculates how many standard errors ($\mathrm{SE}_{\bar{x}}$) away from the null hypothesis μ_0 that the observed statistic $\bar{x}$ is.

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

The t-statistic is used to find the probability of your sample having the mean $\bar{x}$ if the null distribution $\mathrm{N}\left(\mu_0, \mathrm{SE}_{\bar{x}_0}\right)$ were the "true" distribution in your population.

The probability of the sample statistic $\bar{x}$ under the null hypothesis $\mu = \mu_0$ is calculated from the t distribution in R using the function `pt()`. This function takes the test statistic t (called `q` or quantile in R) and the degrees of freedom `df` $(n-1)$ as parameters. This probability is known as the p-value.

```
# left-tailed hypothesis test
pt(-t, df=n-1)

# right-tailed hypothesis test
pt(t, df=n-1, lower.tail=F)

# two-tailed hypothesis test
pt(-t, df=n-1)*2
```

```
pt(t, df=n-1, lower.tail=F)*2
pt(-t, df=n-1)+pt(t, df=n-1, lower.tail=F)
```

A small p-value indicates that the probability of taking a sample of size n with your observed sample mean $\bar{x}$ from the null sampling distribution $\bar{x}_0 \sim \mathrm{N}\left(\mu_0, \mathrm{SE}_{\bar{x}_0}\right)$ is very low.

If the p-value for your t-statistic is less than your significance threshold α (i.e. the Type I Error Rate), this is evidence that the null hypothesis $H_0 \colon \mu = \mu_0$ is not true. Reject the null hypothesis H_0 and accept the alternate hypothesis H_A.

If the p-value for your t-statistic is greater than α, this is not sufficient evidence against the null hypothesis $H_0 \colon \mu = \mu_0$. You fail to reject the null hypothesis H_0.

Paired Means

Difference in Means

Proportions

Table 3: Sample Statistics for Inference of Population Proportions

Measure	Sample Statistic	Population Parameter
Proportion	$\hat{p}$	p
Difference in Proportions	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$

Assumptions

- ***Independence***: sample observations are independent (i.e. random sample).
- ***Sample size***: the sample size should be greater than 20 ($n \geq 20$) with at least 10 successes ($np \geq 10$) and 10 failures ($n(1-p) \geq 10$).
- ***Validity***: sample statistics approximate the population parameters ($\bar{x} \approx \mu$, $s \approx \sigma$)

Single Proportion ($\hat{p}$) - One-Sample Z-test

- A ***1-sample*** Z-test tests if the mean proportion (p) for a population is different from a null value (p_0).
- Sample statistic $\hat{p}$ (proportion) and sample size n are used to infer the sampling distribution of the mean proportion $\hat{p} \sim \mathrm{N}\left(p, SE_{\hat{p}}\right)$.
- Confidence intervals and hypothesis tests for proportions are based on the Z distribution (standard normal).

Confidence Interval

The confidence interval for the mean proportion $\hat{p}$ estimating p is…

$$\hat{p} \pm Z^* \times SE_{\hat{p}}$$

The standard error of $\hat{p}$ is estimated by…

$$\text{SE}_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The critical value from the Z distribution is…

$$Z^* = Z_{\alpha/2} = Z_{1-\alpha/2}$$

A critical value is calculated from the Z distribution in R using the function `qnorm()`. This function takes a probability `p` ($\alpha/2$ or $1-\alpha/2$), a `mean` (default $= 0$), and a standard deviation (`sd`, default $= 1$).

```r
qnorm(alpha/2)

qnorm(1-alpha/2)
```

Hypothesis Test

The null hypothesis of a 1-sample proportion- or Z-test states that the population proportion p is equal to some null value p_0.

$$H_0 : p = p_0$$

The null distribution of $\hat{p}$ given the null hypothesis that $p = p_0$ is $\hat{p} \sim \text{N}\left(p_0, SE_{\hat{p}}\right)$. The standard error of $\hat{p}$ under the null hypothesis $H_0 : p = p_0$ is…

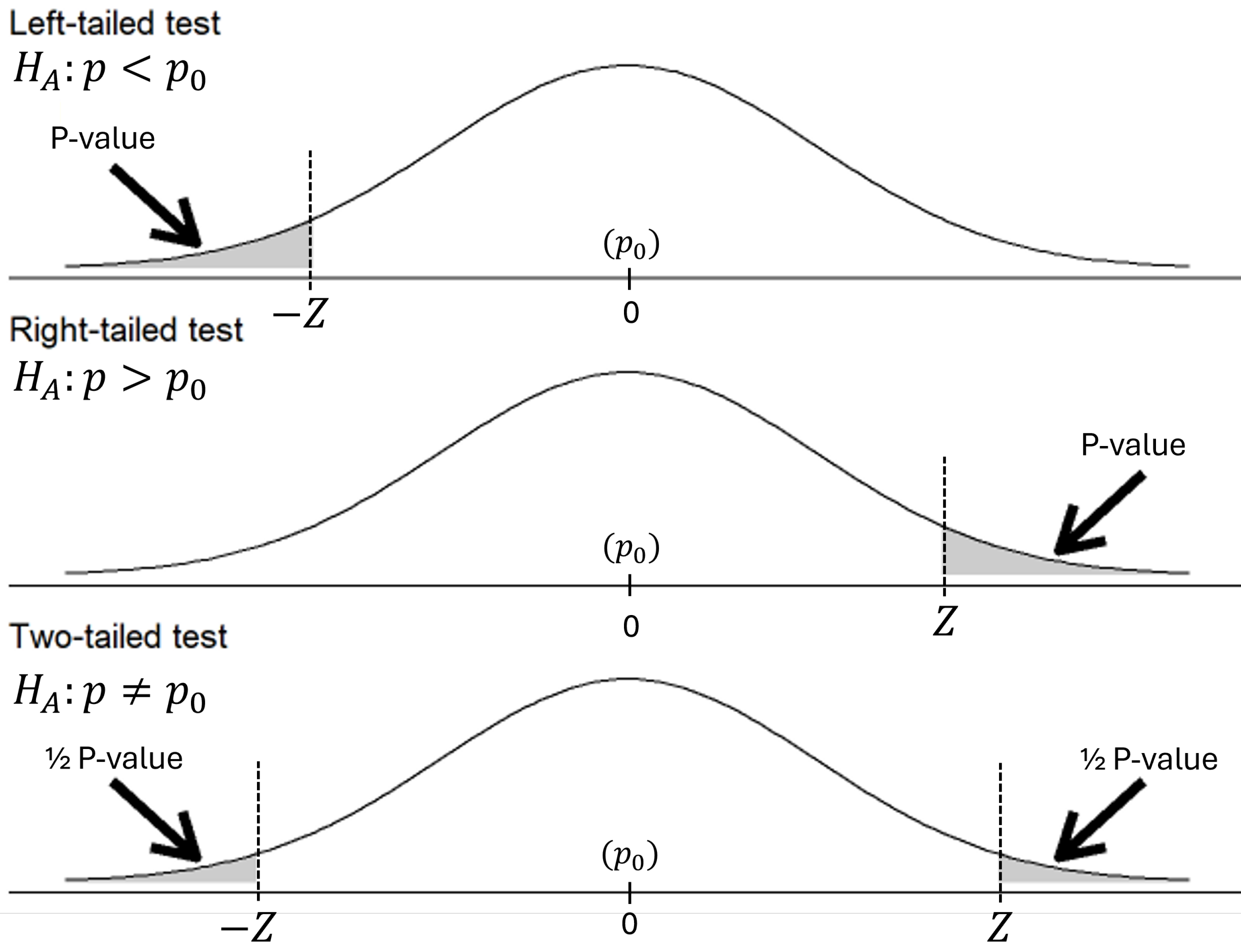
$$\text{SE}_{\hat{p}_0} = \sqrt{\frac{p_0(1-p_0)}{n}}$$

A confidence interval for the mean proportion under the null hypothesis is calculated as before, but using p_0 as the point estimate and $\text{SE}_{\hat{p}}$ under the null hypothesis.

$$p_0 \pm Z^* \times \text{SE}_{\hat{p}_0}$$

The alternate hypothesis of a 1-sample Z-test states that the population proportion p is greater than, less than, or not equal to some null value p_0.

- $H_A : p < p_0$, left-tailed test (one-sided)
- $H_A : p > p_0$, right-tailed test (one-sided)
- $H_A : p \neq p_0$, two-tailed test (two-sided)



The test statistic Z (the Z-statistic) calculates how many standard errors ($\mathrm{SE}_{\hat{p}}$) away from the null hypothesis p_0 that the observed statistic $\hat{p}$ is.

$$Z = \frac{\hat{p} - p_0}{\mathrm{SE}_{\hat{p}_0}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

The Z-statistic is used to find the probability of your sample having the mean proportion $\hat{p}$ if the null distribution $\mathrm{N}\left(p_0, \mathrm{SE}_{\hat{p}_0}\right)$ were the "true" distribution in your population.

The probability of the sample statistic $\hat{p}$ under the null hypothesis $p = p_0$ is calculated from the Z distribution in R using the function `pnorm()`. This function takes the test statistic Z (called `q` or quantile in R), the `mean` (default $= 0$), and the standard deviation (`sd`, default $= 1$) as parameters. This probability is known as the p-value.

```
# left-tailed hypothesis test
pnorm(-Z)

# right-tailed hypothesis test
pnorm(Z, lower.tail=F)

# two-tailed hypothesis test
```

```r
pnorm(-Z)*2
pnorm(Z, lower.tail=F)*2
pnorm(-Z)+pt(Z, lower.tail=F)
```

A small p-value indicates that the probability of taking a sample of size n with your observed sample proportion $\hat{p}$ from the null sampling distribution $\hat{p}_0 \sim \mathrm{N}\left(p_0, \mathrm{SE}_{\hat{p}_0}\right)$ is very low.

If the p-value for your Z-statistic is less than your significance threshold α (i.e. the Type I Error Rate), this is evidence that the null hypothesis $H_0: p = p_0$ is not true. Reject the null hypothesis H_0 and accept the alternate hypothesis H_A.

If the p-value for your Z-statistic is greater than α, this is not sufficient evidence against the null hypothesis $H_0: p = p_0$. You fail to reject the null hypothesis H_0.