Exploratory and Summary Statistics (Chapters 3 & 4)

Statistic	*Parameter*	*Point Estimate*	*Formula*	*Interprétation*	*Notes*
Sum of squares	$\sigma^2 \times df$	SS	$SS = \sum_{i=1}^{n}(x_i - \bar{x})^2$	No easy interpretation.	• Mean and standard deviation are best suited to symmetrical distributions. • When distribution is Normal, 68% of data points lie within $\pm 1\sigma$ of μ, 95% within $\pm 2\sigma$ of μ, and 99.7% lie within $\pm 3\sigma$ of μ • For other distributions, use Chebychev's rule (e.g., at least 75% of data lie within $\pm 2\sigma$ of μ).
Mean	μ	$\bar{x}$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	A measure of central location; balancing pt.	
Variance	σ^2	s^2	$s^2 = \frac{SS}{n-1}$	A measure of spread expressed in units squared	
Standard Deviation	σ	s	$s = \sqrt{s^2}$ or $\sqrt{\frac{SS}{n-1}}$	A measure of spread expressed in data units. More appropriate for descriptive purposes.	

Statistic	*Formula*	*Interpretation*	*5-point Summary*	*Notes of boxplot*
Median	Median has depth of $\frac{n+1}{2}$	A measure of central location	Q0 – Minimum Q1 – First Quartile Q2 – Median Q3 – Third quartile Q4 – Maximum	• Provide information about locations, spread, and shape. The box contains middle 50% of data. Line inside the box is the median. • Anything above the upper fence or below the lower fence is "outside." (Fences are *not* drawn.) Plot outside values as separate points. • The lower whisker is drawn from Q1 to the lower inside value. The upper whisker is drawn from Q3 to the upper inside value.
Interquartile Range (IQR)	$IQR = Q3 - Q1$	A measure of spread, aka "hinge-spread"		
Lower Fence (F_l)	$F_l = Q1 - 1.5(IQR)$	Helps determine: Lower inside value Lower outside value(s)		
Upper Fence (F_u)	$F_u = Q3 + 1.5(IQR)$	Helps determine: Upper inside value Upper outside value(s)		

Probability (Chapters 5–7)

- **Probability** ≡ relative frequency in the population; expected proportion after a very long run of trials; can be used to quantify subjective statements.

- **Properties of probabilities**

  Basic: (1) 0 ≤ Pr(A) ≤ 1; (2) Pr(S) = 1; (3) Pr(Ā) = 1−Pr(A); and (4) Pr(A or B) = Pr(A) + Pr(B) for disjoint events.
  Advanced: (5) If A and B are independent, Pr(A and B) = Pr(A) · Pr(B) (6) Pr(A or B) = Pr(A) + Pr(B) − Pr(A and B) (7) Pr(B|A) = Pr(A and B) / Pr(A) (8) Pr(A and B) = Pr(A) · Pr(B|A) (9) Pr(B) = [Pr(B and A)] + Pr(B and Ā) (10) Bayes' Theorem (p. 111)

- **Binomial variables**: $X \sim \mathrm{b}(n, p)$, $\Pr(X = x) = {}_nC_x p^x q^{n-x}$ where ${}_nC_x = \dfrac{n!}{x!(n-x)!}$ and $q = 1 - p$

- **Cumulative probability:** $\Pr(X \le x)$ = sum all probabilities up to and including $\Pr(X = x)$; corresponds to AUC in the left tail of the *pmf* or *pdf*.

- **Normal variables**: $X \sim N(\mu, \sigma)$. To determine $\Pr(X \le x)$, standardize $z = \dfrac{x - \mu}{\sigma}$ and look up cumulative probability in *Z* table. Use the fact that the AUC sums to 1 to determine probabilities for various ranges.
  To find a value that corresponds to a given probability, look up closest z_p in the Z table and then unstandardize according to $x = \mu + z_p \cdot \sigma$.

Introduction to Inference (Chapters 8–11)

- The **sampling distribution of the mean (SDM)** is governed by the central limit theorem, law of large numbers, and square root law. When *n* is large, $\bar{x} \sim N(\mu, \sigma_{\bar{x}})$ where $\sigma_{\bar{x}}$ is the standard error (*SE*) and is equal to $\dfrac{\sigma}{\sqrt{n}}$. The standard estimate is estimated by $\dfrac{s}{\sqrt{n}}$ when the population standard deviation is not known.
- **(1–α)100% confidence interval for μ.** Use $\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot SE_{\bar{x}}$ when σ is known. Use $\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot SE_{\bar{x}}$ when relying on *s*.
- **Hypothesis testing basics.** Know all the steps, not just the conclusion and keep in mind that hypothesis tests require certain conditions (e.g., Normality, independence, data quality) to be valid. The steps are:
  A. H_0 and H_1 [For one-sample test of a mean, H_0: $\mu = \mu_0$ where μ_0 is the mean specified by the null hypothesis.]
  B. Test statistic [For one-sample test of a mean, use either $z_{\text{stat}} = \dfrac{\bar{x} - \mu_0}{SE_{\bar{x}}}$ or $t_{\text{stat}} = \dfrac{\bar{x} - \mu_0}{SE_{\bar{x}}}$ with $df = n - 1$.]
  C. *P*-value. Convert the test statistic to a *P*-value. Small *P* → strong evidence against H_0.
  D. Significance level. It is unwise to draw too firm a line. However, you can use the conventions regarding marginal significance, significance, and high significance when first learning.
- **Power and sample size basics.** Approach from estimation, testing, or "power" perspective. Sample size requirement for limiting margin of error *m* is given by $n = \left(z_{1-\frac{\alpha}{2}} \dfrac{\sigma}{m}\right)^2$ The power of testing a mean is $1 - \beta = \Phi\left(-z_{1-\frac{\alpha}{2}} + \dfrac{|\Delta|\sqrt{n}}{\sigma}\right)$. The sample size requirement of a one-sample *z* or *t* test: $n = \dfrac{\sigma^2\left(z_{1-\beta} + z_{1-\frac{\alpha}{2}}\right)^2}{\Delta^2}$. It is OK to use *s* as a substitute for σ in power and sample size formulas, when necessary.

Inference

	Parameter* ↕ *estimator	***df***	***Standard Error***	***Confidence Interval***	***Test Statistic***
Chapter 11: Inference about a Mean	μ ↕ $\bar{x}$	$n-1$	$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$	$\bar{x} \pm t_{df,1-\frac{\alpha}{2}} \cdot SE_{\bar{x}}$	To test H_0: $\mu = \mu_0$ $t_{stat} = \frac{\bar{x}-\mu_0}{SE_{\bar{x}}}$
Chapter 12: Comparing Independent Means	$\mu_1 - \mu_2$ ↕ $(\bar{x}_1 - \bar{x}_2)$	df_{Welch} via computer or smaller of df_1 or df_2 for df_{conserv}	$SE_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) \pm t_{df,1-\frac{\alpha}{2}} \cdot SE_{\bar{x}_1-\bar{x}_2}$	To test H_0: $\mu_1 = \mu_2$ $t_{stat} = \frac{\bar{x}_1-\bar{x}_2}{SE_{\bar{x}_1-\bar{x}_2}}$
Chapter 16: Inference About a Proportion	p ↕ $\hat{p}$	N/A	$\tilde{p} \pm z_{1-\frac{\alpha}{2}} \cdot SE_{\tilde{p}}$ where $\tilde{p} = \frac{\tilde{x}}{\tilde{n}}$ where $\tilde{x} = x+2, \tilde{n} = n+4$ and $SE_{\tilde{p}} = \sqrt{\frac{\tilde{p}\tilde{q}}{\tilde{n}}}$ To limit margin of error m, use $n = \frac{z_{1-\alpha/2}^2 \cdot p^* q^*}{m^2}$		To test H_0: $p = p_0$ $z_{stat} = \frac{\hat{p}-p_0}{SE_{\hat{p}}}$ where $SE_{\hat{p}} = \sqrt{\frac{p_0 q_0}{n}}$
Chapter 17: Comparing Two Proportions	$(p_1 - p_2)$ ↕ $(\hat{p}_1 - \hat{p}_2)$	N/A	$(\tilde{p}_1 - \tilde{p}_2) \pm z_{1-\frac{\alpha}{2}} \cdot SE_{\tilde{p}_1-\tilde{p}_2}$ where $\tilde{p}_i = \frac{\tilde{a}_i}{\tilde{n}_i}, \tilde{a}_i = a_i + 1, \tilde{n}_i = n_i + 2$, and $SE_{\tilde{p}_1-\tilde{p}_2} = \sqrt{\frac{\tilde{p}_1\tilde{q}_1}{\tilde{n}_1}+\frac{\tilde{p}_2\tilde{q}_2}{\tilde{n}_2}}$		To test H_0: $p_1 = p_2$ $z_{stat} = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\bar{p}\cdot\bar{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ where $\bar{p} = \frac{a_1+a_2}{n_1+n_2}$