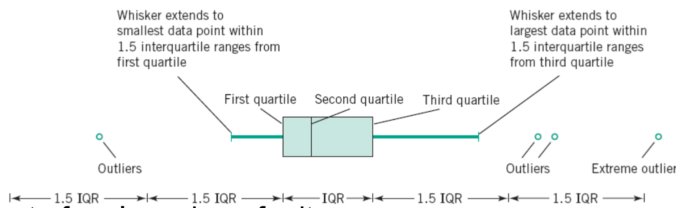
Basic Statistics and Data Visualization

Sample mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, Population variance $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$,

Sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$. Excel function, AVERAGE() for mean and STDEV or (STDEV.S) for s (sample standard deviation), and STDEV.P for σ.

Histogram (distribution); Box Plot

Scatter Diagram shows correlation



Discrete Random Variables

Def: r.v. with a finite (or countably infinite) set of real numbers for its range

Probability mass function (**PMF**): $f(x_i) = P(X = x_i)$;

Cumulative distribution function (**CDF**): $F(x) = P(X \le x) = \sum_{x_i \le x} f(x_i)$

Expectation (or mean) $\mu = E(X) = \sum_{i=1}^{n} x_i f(x_i)$, Variance $\sigma^2 = V(X) = E(X-\mu)^2 = \sum_{i=1}^{n}(x_i-\mu)^2 f(x_i) = \sum_{i=1}^{n} x_i{}^2 f(x_i) - \mu^2$

Bernoulli: $X = 0\ or\ 1$ with $P(X=1) = p$; $E(X) = p$, $V(X) = p(1-p)$

Binomial: $X = 0, 1, 2, \ldots$, representing the number of successes out of n independent trials with the probability of success for each trial being p. $\boldsymbol{f(x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, \ldots, n}$.

$E(X) = np, V(X) = np(1-p)$. *Excel Function BINOMDIST does not work in the online spreadsheet.* Use the equation for calculation.

Poisson: $X = 0, 1, 2, \ldots$, representing the number of events in an interval with rate parameter λ.

$\boldsymbol{f(x) = \frac{e^{-\lambda}\lambda^x}{x!}}$, $E(X) = \lambda$, and $Var(X) = \lambda$. **POISSON**(x, λ, FALSE) for PMF, set to TRUE for CDF.

Continuous Random Variables

Def: r.v. with an interval (either finite or infinite) of real numbers for its range.

Probability density function (**PDF**): $P(a < X < b) = \int_a^b f(x)dx$.

CDF: $(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du$

Mean $\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\, dx$, Variance $\sigma^2 = V(X) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = E(X^2) - \mu^2$

Normal: mean μ, variance σ^2 Standard normal $Z = \frac{X-\mu}{\sigma}$

$P(X \le x) = P(\frac{X-\mu}{\sigma} \le \frac{x-\mu}{\sigma}) = P(Z \le z)$

NORMDIST(x, μ, σ, cumulative) to find probability

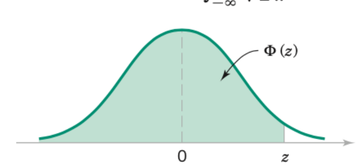
NORMINV(prob, μ, σ) to derive z

Exponential: distance between successive events of a Poisson process with mean $\lambda > 0$

$f(x) = \lambda e^{-\lambda x}$, for $0 \le x < \infty$; $E(X) = \frac{1}{\lambda}$ and $V(X) = \frac{1}{\lambda^2}$

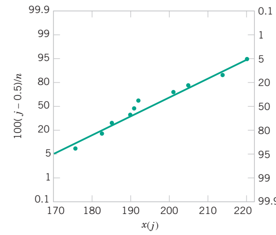
$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} du$$

EXPONDIST(x, λ, cumulative)



Central Limit Theorem: Sample averages will converge to a normal distribution if the sample size is large.

Use **normal probability plot** to check if normality assumption is satisfied.



Statistical Inference

$\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true})$, level of significance

$\beta = P(\text{Type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false})$, Power = $1 - \beta$

The P-value is the smallest level of significance that would lead to the rejection of H_0

7-step Hypothesis Testing Procedure

1. Parameter of Interest; 2. Null hypothesis, H_0; 3. Alternative hypothesis, H_1; 4. Test statistic

5. Reject H_0 if; 6: Computations; 7. Conclusions

One-Sample Mean Test, Variance Known (Z-test)

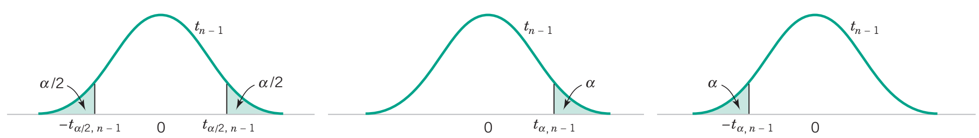
$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

			Confidence interval
$H_1: \mu \neq \mu_0$	Probability above $\lvert z_0\rvert$ and probability below $-\lvert z_0\rvert$, $P = 2[1 - \Phi(\lvert z_0\rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$	$\bar{x} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \le \mu \le \bar{x} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$
$H_1: \mu > \mu_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$	$\bar{x} - z_\alpha\sigma/\sqrt{n} = l \le \mu$
$H_1: \mu < \mu_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$	$\mu \le u = \bar{x} + z_\alpha\sigma/\sqrt{n}$

$$\beta = \Phi\left(z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \qquad n \simeq \frac{(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2} \qquad \delta = \mu - \mu_0$$

One-Sample Mean Test, Variance Unknown (t-test)

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



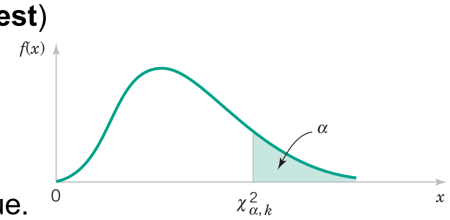
CI: $\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \le \mu \le \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n}$

Use **t.dist** to get probability and use **t.inv** to get t

Inference on the Variance of a Normal Population (Chi-squared test)

$$\chi_0^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

$H_1: \sigma^2 \neq \sigma_0^2$	$\chi_0^2 > \chi^2_{\alpha/2,n-1}$ or $\chi_0^2 < \chi^2_{1-\alpha/2,n-1}$
$H_1: \sigma^2 > \sigma_0^2$	$\chi_0^2 > \chi^2_{\alpha,\,n-1}$
$H_1: \sigma^2 < \sigma_0^2$	$\chi_0^2 < \chi^2_{1-\alpha,\,n-1}$



Use **chisq.dist** to find probability and use **chisq.inv** to find chisq value.

CI: $\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}$

Proportion Test (Z-test)

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1-p_0)}}$$

Follow z-test procedures

CI:

$$\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$